

ALAGAPPA UNIVERSITY

[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle and Graded as Category–I University by MHRD-UGC]
(A State University Established by the Government of Tamilnadu)

KARAIKUDI – 630 003



DIRECTORATE OF DISTANCE EDUCATION

B.Com

IV-SEMESTER

BUSINESS STATISTICS

Author:
Dr. S. NAZEER KHAN
Assistant Professor of Commerce,
PG and Research Department of Commerce,
Dr. Zakir Husain College, Illayankudi

SYLLABI-BOOK MAPPING TABLE

BUSINESS STATISTICS

Syllabi	Mapping in Book
UNIT I Statistics - Importance of Statistics in modern business environment - Definition of Statistics - Scope and Applications of Statistics Characteristics of Statistics- Functions of Statistics - Limitations of Statistics - Statistical Software	**Pages - 1-15**
UNIT II Measures of Central Tendency and Dispersion - Objectives of statistical average - Requisites of a Good Average - Statistical Averages - Arithmetic mean - Properties of arithmetic mean - Merits and demerits of arithmetic mean \- Median - Merits and demerits of median - Mode - Merits and demerits of mode - Geometric Mean-Harmonic Mean - Appropriate Situations for the Use of Various Averages - Positional Averages - Dispersion – Range - Quartile deviations - Mean deviation, Standard Deviation - Properties of standard deviation Coefficient of Variance..	**Pages - 16-58**
UNIT III Theory of Probability - Definition of probability - Basic terminology used in probability theory - Approaches to probability - Rules of Probability - Addition rule - Multiplication rule - Conditional Probability - Steps Involved in Solving Problems on Probability - Bayes' Probability - Random Variables	**Pages - 59-75**
UNIT IV Theoretical Probability Distributions - Random variables - Probability Distributions - Discrete probability distributions - Continuous probability distributions - Bernoulli Distribution - Repetition of a Bernoulli experiment - Binomial Distribution - Assumptions for applying a binomial distribution - Examples of binomial variate - Recurrence formula in case of binomial distribution - Case study on binomial distribution Poisson Distribution - Assumptions for applying the Poisson distribution - Real life examples of Poisson variate - Recurrence relation -Case study on Poisson distribution - Normal Distribution - Standard Normal Distribution	**Pages – 76-95**
UNIT V Estimation: Reasons for Making Estimates - Making Statistical Inference - Types of Estimates - Point estimate - Interval estimate - Criteria of a Good Estimator - Unbiasedness - Efficiency - Consistency – Sufficiency - Point Estimates – Interval Estimates - Case study on calculating estimates - Making the interval estimate Interval Estimates and Confidence Intervals - Interval estimates of the mean of large samples - Interval estimates of the proportion of	**Pages - 96-103**

CONTENTS

BLOCK III: CORRELATION, REGRESSION, BUSINESS FORECASTING

BLOCK I: FUNDAMENTALS OF STATISTICS

UNIT I - STATISTICS

Structure

1.0 INTRODUCTION

Statistics is an area of study that deals with collecting, organising, analyzing, interpreting and presenting data. The study of statistics has a lot of applications in industries, agriculture, medicine etc. In this unit, you will learn about the various importance and scope of statistics. You will also come to know about types of data, ways of collecting them and also how to present data.

1.1 OBJECTIVES

After going through this unit, you will

- Understand the meaning , importance and functions of statistics
- Learn the different types of data and how to collect them.
- Know how the data collected can be presented

1.2 STATISTICS

The word statistics of English language have been derived from the Latin word status or Italian word ‘statista’ or German word ‘statistik’. In each case it means "an organised political state”. Although, in the past, statistics was considered as the "science of statecraft" as it was used by the government of various States to collect data regarding population ,births , deaths, taxes etc.,. Statistics, nowadays, have experienced a modern development. Statistics play a crucial role in enriching a specific domain by collecting data in that field, analyse the data by applying various statistical techniques and making inferences about the same.For example, knowing the average height of the students will enable the engineer to know about the size of the door.

1.2.1 DEFINITION OF STATISTICS

The definition of statistics can be expressed in two ways to cover two different concepts. They are

1. Statistics as numerical data
2. Statistics for statistical method

1. Statistics as numerical data

When the word ‘statistics’ is used in plural sense, it refers to the collection of numerical data.

For example: - Export or Import quantity, Foreign Direct Investment, etc..,.

According to **Webster**,” statistics are classified facts representing the conditions of the people in a state especially those facts which can be stated in number or in table of numbers or in any tabular or classified arrangements"

This definition of Webster reveals that only numerical facts can be termed statistics. This is an old, narrow and inadequate definition for modern times.

According to **Bawley** “Statistics are numerical statement of facts in any department of inquiry placed relation to each other"

Here, Bowley says that statistics is the science of counting and ignores other aspects such as analysis, interpretations etc..,.

According to **Yule and Kendall**,” By statistics we mean quantitative data affected to a market extent by multiplicity of cause"

Yule and Kendall’s definition tells us that numerical data is affected by multiplicity of cause. For example, the cost of production is affected by wage cost, exchange rate, raw material etc..,.

According to **Professor Horace Secrist**," It is the aggregate of facts affected to mark extent by multiplicity of causes, numerically

expressed, enumerated or estimated according to a reasonable standard of accuracy, connecting in a systematic manner for the predetermined purpose and placed in relation to each other"

Secrist's definition for statistics is more complete. The vital point that the definition covers are

1) Aggregate of facts
2) Affected by multiplicity of cause
3) Numerically expressed
4) Estimated according to standard of accuracy
5) Systematic Collection of data
6) Data collected for a predetermined purpose
7) Comparable

2. Statistics as Statistical Methods

According to **Bowley**," Statistics the science of measurement of social organism, regarded as a whole in all its manifestation"

This definition of Bowley is insufficient

According to **Wallis and Roberts**," Statistics is a body of methods for making wise decision on the face of uncertainty"

This definition is modern as it conveys statistical methods enable us to arrive at valid decisions.

According to **Croxton and Cowden**" statistics must be defined as the science of collection, presentation, analysis and interpretation of numerical data"

This definition gives a more elaborate meaning to statistics as statistical tools.

1.2.2 IMPORTANCE OF STATISTICS

Statistics can be used to various areas of business operations for effective results. Some prominent areas are given below.

1) **Startups -** While opening a new business or acquire one, we need to study the market from a statistical point of view to get accuracy in the market demand and supply .A businessman must do proper research by collecting data, analyzing and interpreting them regarding market trends before starting his business.
2) **Production -** The production of the commodity depends upon various factors such as demand, supply of capital etc..,. These factors must be analyzed statistically to get a precise and accurate view of the same.
3) **Marketing -** An ideal marketing strategy requires statistical analysis on population, income of consumers, availability of the product ect..,.

4) **Investment -** Statistics play a vital role in making decisions regarding buying shares, debentures or real estate. Using this statistical data, an investor will buy investments at a lesser price and sell when the price increases.
5) **Banking -** Banking sector is highly influenced by economic and market conditions. Bank have separate research department which collect and analyse information regarding inflation rate, interest rates, bank rates etc..,.

1.2.3 LIMITATIONS OF STATISTICS

1) Statistics does not analyse qualitative phenomenon

As statistics is a science which deals with numerical, it cannot be applied in data that cannot be measured in terms of quantitative measurements. However statistical techniques can be used to convert the qualitative data to quantitative data.

2) Statistics does study individuals

Statistics deals with aggregate quantities and doesn't give importance to individual data. This is because individual data is not useful for statistical analysis.

3) Statistical laws are not exact

Statistical interpretations are based on averages and hence are only approximations can be made

4) Statistics may be misused

Statistical data when used by an inexperienced person or illiterate person can lead to wrong interpretations. Hence it must be used only by experts.

1.2.4 FUNCTIONS OF STATISTICS

1) Consolidation

Statistics enables you to consolidate and understand huge data by providing only significant observations.

For example, instead of observing the marks of each and every individual with class average will enable you to know the class's performance as a whole.

2) Comparison

Classification and tabulation of data are used to compare the data. Various statistical tools such as graph, measure of depression dispersion, correlation gives us huge scope for comparison.

For example, the market demand for a product can be compared among the states. This enables the company to identify and analyse the target market.

3) Forecasting

Forecasting means predicting the future prospects. Statistics plays a huge role in forecasting the future.

For example, with the data of the sales value for the past 10 years, we will be able to predict the sales of the coming year approximately. Time series analysis and regression analysis are important for forecasting.

4) Estimation

One of the main aims of statistics is to draw conclusions on a huge population based on the analysis from a sample group.

For example, from a sample height of 10 students will be able to estimate the average height of all the students from the class.

5) Test of hypothesis

Statistical hypothesis is portraying a huge population from the inferences of a sample observation.

For example, if a particular fertilizer helps in increasing the crop yield in a particular area then it will be used in other areas based on this sample.

1.2.5 SCOPE OF STATISTICS

1) Statistics in Industries

Statistics is extensively used in huge number of industries. Statistics may be used in sales forecasting, consumer preference, quality control, inventory control, risk management etc. Sampling is vital for inspection plans.

2) Statistics in Education

Statistics plays an important role in education. Statistics help in measuring and evaluating the progress of the student, formulating policies and also helps to predict the future performance of the students to help them improve in the same.

3) Statistics in Economics

Statistics helps us to understand and analyse economic theories. Right from analysing microeconomic factors like the demand for the product, research regarding different markets to macroeconomic concept like inflation, unemployment can be done easily using statistics.

4) Statistics in Medicine

Statistics helps in researching and analysing medical experiments and investigations. Biostatic enables researchers to identify if a particular treatment or drug is working and how effective it is.

5) Statistics in Modern Application

A lot of software's are developed day to day for experimentation, forecasting and estimation.

For example, SYSAT is one such software which provides with scientific and technical graphical options.

6) Statistics in Agriculture

Statistics can be applied in agriculture by analysing the effectiveness of fertilizers. It can be used in taking decisions regarding inputs and outputs, inventories etc..,.

1.3 DATA

Data are pieces of factual information that are recorded and applied for analysis. Data is a tool which helps us to understand certain problems by providing us with information. They are a set of values with qualitative and quantitative variable.

1.3.1 TYPES OF DATA

Data of broadly classified into two based upon who collected the data

Primary data

Primary data is the data collected by investigator himself for the first time for his own research and analysis. It is also known as first-hand information. Primary data is collected using method such as personal interview, survey etc..,.

Secondary data

Secondary data is the data which is already been collected and process by the person for the purpose of his research. Journals, internal sources, journals, book etc..,. are sources of secondary data.

CHECK YOUR PROGRESS -1

1. What are the two ways in which statistics can be defined?
2. What is the definition of statistics according to Professor Horace Secrist?
3. How does statistics help in comparing data?
4. What is the role of statistics in medicine?
5. What is secondary data?

1.4 DATA COLLECTING TECHNIQUES

1.4.1PRIMARY DATA

1) Direct Personal Investigation

Direct personal investigation is the method in which the investigator directly goes to the source to collect information.

Merits

(i) Information collected in this method is more authentic and accurate
(ii) There is high degree of accuracy in qualitative information
(iii) The original opinion or data shall be obtained.

Demerits

(i) This is a time consuming process
(ii) If the investigator is not intelligent enough to understand the mental state of the source it may lead to wrong interpretation.
(iii) It may result in personal bias.

2) Indirect Oral Investigation

Indirect oral investigation is when the investigator investigates a person close to the source. This is done due to the reluctance of the original person.

Merits

(i) It saves time and labour
(ii) It is easy and convenient
(iii) It covers a wide range of area.

Demerits

(i) Information received may not be reliable
(ii) Person chosen for this purpose me not be suitable
(iii) It may be expensive as information is collected from various sources.

3) Information collected from local agencies

In this method investigator appoints a few agencies in various regions to cover various fields of inquiry. This method is generally used by newspaper companies to get information from various places in various topics such as sports, economics etc..,.

Merits

(i) Avoid area can be easily covered
(ii) This is a time saving method of collecting data
(iii) The cost of collecting data is less

Demerits

(i) Sometimes the information collected may contradict one another
(ii) The information can be less accurate
(iii) This method will be expensive and a full-time agent is hired in different places

4) Questionnaire method

Questionnaire method is the most famous method of collecting primary data .A questionnaire is a set of questions device for conducting survey. The questionnaire is sent to the respondent with the request to fill it and send it back within a specific time.

Merits

(i) This method is cheaper
(ii) The time consumed for this process is very less
(iii) This is an unbiased method of collecting data

Demerits

(i) Sometimes the respondent may provide wrong information
(ii) There is no type of personal motivation in this method
(iii) There are chances of ignorance or late reply from the respondents

General principles of framing a questionnaire

1) The questionnaire must not be very long

We must try to give the questions as minimum as possible. Long questionnaire may lead to boredom or discontentment among the respondents.

2) The question must move from general to specific

When the question moves from general to specific respondent become more comfortable in answering the questions

3) The question should be ambiguous

The questions must be in such ways that the respondents are able to give clear and quick answers to the questions

4) The person should not contain double negatives

Words like don't you or wouldn't you must not be used in the questions as they might tempt the respondent to give a biased answer.

5) The question should not be lending questions

The questions should not give clues to the respondent on how they must answer it.

6) The question must not provide alternators for the answer

For example, instead of asking would you like to do engineering or medicine after class 12, the correct way of asking the question is would you like to do engineering?

1.4.2 SECONDARY DATA

1) Published sources

Certain government and non-government organisations publish various journals, research papers, surveys etc which are very helpful and reliable. Some of them are mentioned below

(i) Publications of international bodies like UNO, WTO and WHO etc..,.
(ii) Publications of research institutes like ISI, NCERT, ICAR etc..,.
(iii) Government publications
(iv) Publications of commercial and financial institutions
(v) Publications of governmental organisations
(vi) Newspaper, journals and periodicals.

2) Unpublished sources

Unpublished sources cover all the sources where data is maintained privately by certain private agencies or companies. The data collected by universities, research institutions also come under unpublished sources.

1.5 PRESENTATION OF DATA

In the previous topic we saw how data can be collected .As the data collected is generally huge we need to comprise and deliver it in a presentable form. Generally there are three ways of presenting presentation of data. They are

1) Textual or Descriptive Presentation
2) Tabular Presentation
3) Diagrammatic Presentation

1.5.1 Textual or Descriptive Presentation

When the data collected is presented in the form of a text it is called textual or descriptive presentation. Generally this method cannot be used to present large data.

For example, in the 2011 census, the population of India was 1,21,08,54,977 comprising of 58, 64, 69,174 females and 62, 37, 24,248 males. The literacy rate is 74.0 4 percentage and density of population is 382 person per square kilometer.

From the above example, we can see that the data is represented textually. One of the major limitations of this method is that the readers must go through the entire text and get the required information.

1.5.2 Tabular Presentation of Data

When the data is presented in the form of rows and columns it is called tabular presentation of data.

Example:

AREA	FEMALE	MALE	TOTAL
URBAN	90%	89%	89.5%
RURAL	87%	88%	87.5%
TOTAL	88.5%	88.5%	88.5%

The about table represents the pass percentage of the examination conducted in Tamilnadu it has three rows (urban, rural, total) and three columns (female, male, total). It is a 3×3 table where each small box is called the cell which gives information regarding the pass percentage. This method is very significant as it enables us to use it for further statistical treatment. This tabular representation is further classified into four

(i) Qualitative Classification

Qualitative classification is when the collected information is classified in the form of attributes such as gender, nationality etc..,. The table given above is an example of qualitative classification where the information is classified in the form of gender and location.

(ii) Quantitative Classification

When information can be measured quantitatively like age, income, marks etc..,.then, such classifications are called quantitative classification

Example

MARKS	FREQUENCY
0-10	5
10-20	10
20-30	20
30-40	15
40-50	10

(iii) Temporal Classification

Temporal classification is when classification is based on the basis of time like year, months, days etc..,.

Example

DAYS OF A WEEK	PRODUCTION (no of pairs of shoes)
MONDAY	2000
TUESDAY	1750
WEDNESDAY	3000
THURSDAY	2250
FRIDAY	1550

(iv) Spatial Classification

Spatial classification is when the data classification is based on place like town, city, district, state, country etc..,.

Example

STATE	LITERACY RATE
TAMIL NADU	80.09%
ANDHRA PRADESH	67.02%
KARNATAKA	75.36%
KERALA	93.91%

1.5.3 Diagrammatic Presentation

In this method the data is represented diagrammatically and is very easy to understand generally data is represented diagrammatically in three ways.

1) Geometric Diagram

This category consists of bar diagrams and pie charts

(i) Bar diagram

Bar diagram is a diagrammatic representation of data in equal spaced and equalwidth rectangular bars for each class of data .The height or length of the bar tells us about the magnitude of the class. Bar diagrams can be easily used for comparison of data. Both qualitative and quantitative data can be represented in bar diagram.

They can be further divided into two broad categories.

a) Multiple bar diagram

When there is a need to compare two set of data multiple bar diagram is used. For example import and export, production and sale etc..,.

b) Component bar diagram

Component bar diagram also known as Sub diagrams are used to compare different components of a particular class. For example, the various components such as rent, medicine, education on which the monthly salary spend can be easily understood from a component bar diagram.

(ii) Pie diagram

A pie diagram is similar to that of a component bar diagram but it is represented in circle proportionally instead of bars. The values given in each class is converted into percentage and then each figure is multiplied by 3.6 degree. (360/100 - 360 degree of a circle divided into 100 parts) the values are then divided accordingly in the circle.

2) Frequency diagram

When the data is in the form of grouped frequency are usually represented by frequency diagrams. Histogram, frequency polygon, frequency curve and ogive are types of frequency diagram.

(i) Histogram

Histogram is a diagram which consists of rectangular bars whose area is proportional to the frequency of a variable and whose width is equal to the class interval.

(ii) Frequency polygon

A frequency polygon is another type of frequency distribution graph. In a frequency polygon, the number of observations is marked with a single point at the midpoint of each and every interval. Then the points are connected using a straight line.

(iii)Frequency curve

The frequency curve is obtained by drawing a smooth freehand curve that passes through the points of a frequency polygon closely as possible.

(iv)Ogive

Ogive also known as the cumulative frequencies are of two types. When the cumulative frequencies are plotted against their upper limits respectively, then it is less than ogive. When the

cumulative frequencies are plotted against their lower limits respectively, then it is more than ogive.

3) Arithmetic line graph

An arithmetic line graph also known as time series graph is a graph where the time (months, years, weeks) are plotted in the x axis and their respective values are plotted in the y axis. It helps us in analysing trends and periodicity of data.

CHECK YOUR PROGESS - 2

6. What is indirect oral investigation?
7. State two merits of questionnaire method
8. Give some examples of published sources
9. What is component bar graph?
10. What is spatial classification?

1.6 SUMMARY

- The word 'statistics' is used in plural sense refers to the collection of numerical data and in singular sense it means the science of collecting, classifying and using statistics
- Statistics can be used to various areas of business operations such as start-ups, production, and marketing for effective results.
- Data is a tool which helps us to understand certain problems by providing us with information. It can be further divided into primary and secondary data.
- Direct personal investigation, indirect oral investigation, questionnaire methods are some of the methods of collecting primary data. Publications of international bodies, research institutions are methods of collecting secondary data.
- Data can be presented in three ways. They are Textual or descriptive presentation, Tabular presentation, Diagrammatic presentation.

1.7 KEY WORDS

Statistics, data, Primary data, Secondary data, Direct personal interview, Indirect oral investigation, Questionnaire, Qualitative, Quantitative, Temporal, Spatial, Bar diagram, Pie diagram, Histogram, Frequency Polygon, Frequency curve, Ogive , Arithmetic line graph.

1.8 ANSWERS TO CHECK YOUR PROGRESS

1. The word 'statistics' is used in plural sense refers to the collection of numerical data and when in singular sense it means the science of collecting, classifying and using statistics
2. According to Professor Horace Secrist," It is the aggregate of facts affected to mark extent by multiplicity of causes, numerically expressed, enumerated or estimated according to a reasonable standard of accuracy, connecting in a systematic manner for the predetermined purpose and placed in relation to each other
3. Classification and tabulation of data are used to compare the data. Various statistical tools such as graph, measure of depression dispersion, correlation gives us huge scope for comparison.
4. Statistics helps in researching and analysing medical experiments and investigations. Biostatic enables researchers to identify if a particular treatment or drug is working and how effective it is.
5. Secondary data is the data which is already been collected and process by the person for the purpose of his research.
6. Indirect oral investigation is when the investigator investigates a person close to the source. This is done due to the reluctance of the original person.
7. Questionnaire method
   (i) This method is cheaper
   (ii) The time consumed for this process is very less.
8. Publications of international bodies like UNO, WTO and WHO, Publications of research institutes like ISI, NCERT, ICAR, and Government publications.
9. Component bar diagram also known as Sub diagrams are used to compare different components of a particular class.
10. Spatial classification is when the data classification is based on place like town, city, district, state, country etc..,.

1.9 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. Write short notes about the types of date
2. List the merits and demerits of direct personal interview
3. What are the general principles followed while framing a questionnaire?
4. Write about the classification of tabular presentation of data.
5. What is a bar diagram? What are its types?

LONG ANSWER QUESTIONS

1. Analyse the importance and scope of statistics
2. Explain in detail about the data collection techniques used in

primary data.

3. Discuss about the functions and limitations of statistics.
4. Explain the various methods used for presentation of data.

1.10 FURTHER READINGS

1. Gupta, S. P. : Statistical Methods, Sultan Chand and Sons, New Delhi.
2. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.
3. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall,NJ.
4. Levin, Richard I. and David S. Rubin: Statistics for Management, Prentice Hall, New Delhi.
5. Lawrance B. Moore: Statistics for Business & Economics, Harper Collins, NY.
6. Watsman Terry J. and Keith Parramor: Quantitative Methods in Finance International, Thompson Business Press, London.

UNIT II MEASURES OF CENTRALTENDENCY

Structure

2.0 INTRODUCTION

Measures of central tendency are a statistical tool used to summarize data that depicts the central value of the given data. These measures enable us to identify where most of the values fall. The three most commonly used measures of central tendency are mean, median and mode. In this unit you will learn about them extensively and also learn about some other partition values.

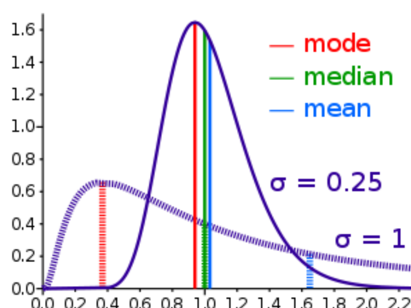
2.1 OBJECTIVES

From this unit you will

- Learn about the measures of central tendency
- Come to know about the various methods of calculating mean, median and mode.
- Know about the partition values.

2.2 MEASURES OF CENTRAL TENDENCY

When working on a given set of data, it is not possible to remember all the values in that set. But we require inference of the data given to us. This problem is solved by mean, median and mode. Measures of Central Tendency, represent all the values of the data. As a result, they help us to draw an inference and an estimate of all the values. They are also known as statistical averages. Their simple function is to mathematically represent all the values in a particular set of data. Hence, this representation shows the general trend and inclination of all the values.



An average provides a simple way of representation of all the individual data. It also aids in the comparison of different groups of data. In addition to this, an average in economic terms can represent the direction an economy is headed towards. Hence, it can be easily used to formulate policies and bring about a reform for a better economy.

2.3 MEAN

2.3.1 ARITHMETIC MEAN

The arithmetic mean of a series of numbers is sum of all observations divided by the total number of observations in the series.

Example:

There are two brothers, with different heights. The height of the younger brother is 138 cm and height of the elder brother is 154cm. The average height of the two brother is total height divided into two equal parts,

$$(138+154) \div 2 = 292 \div 2 = 146 \text{ cm}$$

So 146 cm is the average height of the brothers. Here 154 > 146 > 138. The average value lies in between the minimum value and the maximum value.

Thus if x1, x2, ..., x_n represent the values of n observations, then arithmetic mean (A.M.) for n observations is: (direct method)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

There are two methods for computing the arithmetic mean: (i) Direct method (ii) Short cut method.

Direct Method:

Example:

The following data represent the number of books issued in a college library is selected from 7 different days 17,1 9, 22, 25, 15, 40, 21 find the mean number of books.

Solution:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{x} = \frac{20 + 39 + 22 + 25 + 45 + 40 + 54}{7} = \frac{245}{7} = 35$$

Hence the mean of the number of books is 35

Indirect Method:

In this method an assumed mean or an arbitrary value (A) is used as the basis of calculation of deviations (d_i) from individual values. If $d_i = x_i - A$

$$\overline{x} = A + \frac{\sum_{i=1}^{n} d_i}{n}$$

Example:

A student's marks in 5 subjects are 95, 78, 88, 72,99. Find the average of his marks.

Let us take the assumed mean, A = 88

x_i	$d_i = x_i - 88$
95	7
78	10
88	0

72	-16
99	10
Total	11

Solution:

$$\bar{x} = A + \frac{\sum_{i=1}^{n} d_i}{n}$$

$$= 88 + \frac{11}{5} = 88 + 5.5 = 93.5$$

The arithmetic mean of average marks is 93.5

Discrete Grouped data

If x1, x2, ...,xn are discrete values with the corresponding frequencies f1, f2, …, fn.

Then the mean for discrete grouped data is defined as (direct method)

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

In the short cut method the formula is modified as

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} \quad \text{where} \quad d_i = x_i - A$$

Example:

Given the following frequency distribution, calculate the arithmetic mean

Marks	64	63	62	61	60	59
No. Of. Students	8	18	12	9	7	6

Solution:

x_i	f_i	$f_i\, x_i$	$d_i = x_i - A$ (A=62)	$f_i d_i$
64	8	512	2	16
63	18	1134	1	18
62	12	744	0	0
61	9	549	-1	-9
60	7	420	-2	-14
59	6	354	-3	-18
	60	3713		-7

Direct Method

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

$\bar{x}$ = 3713 / 60 = 61.88

Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N}\ x\ c$$

Here A = 62

$\bar{x} = 62 - \frac{7}{60} = 61.88$

The mean mark is 61.88

Mean of continuous Grouped data:

Direct method

$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$, x_i is the midpoint of the class interval

Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N}\ x\ c$$

$$d = \frac{x_i \; - A}{c}$$

Where A – any arbitrary value

c - Width of the class interval

x_i -is the midpoint of the class interval

Example:

For the frequency distribution of yield of tomato given in table calculate the mean yield per plot.

Yield per plot (in Kg)	64 - 84	84 - 104	104 – 124	124 – 144
No of plots	3	5	7	20

Solution:

Yield (in Kg)	**No of plots (f_i)**	**Mid x_i**	**f_i x_i**	**d = (x_i - A) / c**	**$f_i d_i$**
64 - 84	3	74	222	-1	-3
84 - 104	5	94	470	0	0
104 – 124	7	114	798	1	7
124 – 144	20	134	2680	2	40
Total	**35**		**4170**		**44**

Direct Method

$$\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{N}$$

$\bar{x}$ = 4170 / 35 = 119.143

Short cut method

$$\bar{x} = A + \frac{\sum_{i=1}^{n} f_i d_i}{N} \; x \; c$$

$$\bar{x} = 94 + \frac{44}{35} \text{ x c } = 119.143$$

3.3.2 WEIGHTED ARITHMETIC MEAN

For calculating simple mean, all the values or the sizes of items in the distribution have equal importance. But in practical life this may not be so, in case some items are more important than others, a simple average computed is not representative of the distribution. Proper weightage has to be given to the various items.

For example a student may use a weighted in order to calculate their percentage grade in a course, in this the student would multiply the weighing of all assessment items in the course(eg: assignment, exams,

projects, etc.)by respective grade that was obtained in each of categories

The average whose component items are being multiplied by certain values known as "weights" and the aggregate of the multiplied results are divided by the total sum of their "weight"

Let $x_1, x_2, \ldots, x_n$ be the set of n values having weights w1,w2,....,wn respectively,

then the weighted mean is

$$\bar{x}_{w} = \frac{w_1\ x_1\ +\ w_2\ x_2 +\ \ldots\ldots\ldots w_n\ x_n}{w_1\ +\ w_2\ +\ w_{3+\cdots\ldots\ldots+}w_1} = \frac{\sum_{i=1}^{n} w_i\ x_i}{\sum_{i=1}^{n} w_i}$$

Example:

A student obtained the marks 40,50,60,80, and 45 in math, statistics, physics, chemistry and biology respectively. Assuming weights 5,2,4,3, and 1 respectively for the above mentioned subjects, find the weighted arithmetic mean per subject.

Solution

Components	Marks scored (x_i)	Weightage (w_i)	$w_i\ x_i$
Maths	40	5	200
Statistics	50	2	100
Physics	60	4	240
Chemistry	80	3	240
Biology	45	1	45
Total		**15**	**825**

Weighted average:

$$\overline{x} = \frac{\sum w_i x_i}{\sum w_i}$$

= 825 / 15 = 55 marks / subject

Combined Mean:

In the arithmetic averages and the number of items in two or more related groups are known, the combined or the composite mean of the entire group can be obtained by

$$\bar{x}_{12} = \frac{n_1 \overline{x_1} + n_2 \overline{x_2}}{n_1 + n_2}$$

The advantage of combined arithmetic mean is that we can

determine the overall mean of the combined data without going back to the original data

Example:

If a sample size of 22 items has a mean of 15 and another sample size of 18 items has a mean of 20. Find the mean of the combined sample?

Solution:

$$\bar{x}_{12} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

$$= \frac{22 \times 15 + 18 \times 20}{22 + 18}$$

$$= \frac{330 + 360}{40} = \frac{690}{40} = 172.5$$

Merits of AM

1. It can be calculated easily and is also easy to understand.
2. Fluctuation can be minimized
3. It can further be used for statistical treatement like median,mode etc.,.
4. This method is rigidly defined and hence can be used for comparison

Demerits of AM

1. It cannot be plotted in a graph.
2. It is not applicable in qualitative data.
3. AM cannot be calculated if the class intervals have open ends.
4. It is highly influenced by extreme observations.

3.3.2 GEOMETRIC MEAN (GM)

A geometric mean is a mean or average which shows the central tendency of a set of numbers by using the product of their values.

The geometric mean of two numbers, say x, and y is the square root of their product x×y. For three numbers, it will be the cube root of their products i.e., (x y z) 1⁄3.

The geometric mean of a series containing n observations is the nth root of the product of the values. If x1, x2,……xn are observations then

$$\text{G. M.} = \sqrt[n]{x_1 . x_2 ... x_n}$$

$$= (x_1 . x_2 ... x_n)^{\frac{1}{n}}$$

$$\log \text{G.M.} = \log (x_1 . x_2 . . . x_n)$$

$$= (\log x_1 + \log x_2 + . . . + \log x_n)$$

$$= \frac{\sum_{i=1}^{n} \log x_i}{n}$$

$$\text{G.M.} = \text{Antilog} \frac{\sum_{i=1}^{n} \log x_i}{n}$$

Example:

Calculate the geometric mean of the following growth of price of onions per 100 Kg per annum is 180, 250, 490, 1400, and 1050

x	180	250	490	1400	1050	**Total**
log x	2.2553	2.3979	2.6902	3.1461	3.0212	**13.5107**

Solution:

$$\text{G.M.} = \text{Antilog} \frac{\sum_{i=1}^{n} \log x_i}{n}$$

$$= \text{Antilog } \frac{13.5107}{5}$$

$$= \text{Antilog } 2.7021 \quad = 503.6$$

Geometrical mean of onion rate is 503.6

Example:

Find the geometric mean for the following distribution of student's marks:

Marks	0 – 30	30 – 50	50 – 80	80 - 100
No . of students	20	30	40	10

Solution:

Marks	No of students f	Mid points x	f log x

0 – 30	20	15	20 (log 15) = 20(1.1761) = 23.5218
30 – 50	30	40	30 (log 40) = 30 (1.6020) = 48.0168
50 – 80	40	65	40 (log 65) = 20(1.8129) = 72.5165
80 - 100	10	90	10 (log 90) = 20(1.9542) = 19.5424
Total	**100**		**163.6425**

$$\text{G.M.} = \text{Antilog}\frac{\sum_{i=l}^{n}\log x_i}{n}$$

$$= \text{Antilog}\ \frac{163.6425}{100}$$

$$= \text{Antilog}\ 1.6364 \quad = 503.6$$

Geometrical mean of onion rate is 43.29

Merits of Geometric mean:

1. It is strictly defined
2. It is based on all items
3. It is very suitable for averaging ratio, rates and percentages
4. It is capable of further mathematical treatment
5. Unlike AM, it is not affected much by the presence of extreme values

Demerits of geometric mean:

1. It cannot be used when the values are negative or if any of the observations is zero
2. It is difficult to calculate particularly when the items are very large or when there is a frequency distribution
3. It brings out the property of the ratio of the change and not the absolute difference of change as the case in arithmetic mean
4. The GM may not be the actual value of the series

3.3.3 HARMONIC MEAN

Harmonic mean of a set of observations is defined as the reciprocal of the arithmetic average of the reciprocal of the given values. If $x_1,x_2.....x_n$ are n observations.

A harmonic mean is used in averaging of ratios. The most common examples of ratios are that of speed and time, cost and unit of material, work and time etc. The harmonic mean (H.M.) of n

observations is

H.M. for ungrouped data

$$\text{H. M.} = \frac{n}{\sum_{i=l}^{n}\left(\frac{1}{x_i}\right)}$$

Example:

Calculate the harmonic mean of the numbers 13.5, 14.5, 14.8, 15.2 and 16.1

Solution:

The harmonic mean is calculated as below:

x	1 / x
13.2	0.0758
14.2	0.0704
14.8	0.0676
15.2	0.0658
16.1	0.0621
Total	**0.3417**

$$\text{H. M.} = \frac{n}{\Sigma\left(\frac{1}{x_i}\right)}$$

$$= \frac{5}{0.3417} = 14.63$$

H.M. Discrete Grouped data:

For a frequency distribution

$$\text{H. M.} = \frac{N}{\sum_{i=l}^{n} f_i\left(\frac{1}{x_i}\right)}$$

Example:

The frequency distribution of first year students of a particular college, calculate the harmonic mean

Age (years)	17	18	19	20	21
	2	5	13	7	3

Solution:

Age (years) x	Number of students f	f / x
17	2	0.1176
18	5	0.2778
19	13	0.6842
20	7	0.3500
21	3	0.1429
Total	**30**	**1.5725**

$$\text{H. M.} = \frac{N}{\sum_{i=1}^{n} f_i\left(\frac{1}{x_i}\right)}$$

$$= 30 / 1.5725 = 19.0779 \approx 19 \text{ years}$$

Merits of H.M:

1. It is strictly defined
2. It is defined on all observations.
3. It is amenable to further algebraic actions
4. It is most suitable average when it is desired to give greater weight to smaller observations and less weight to larger observations.

Demerits of H.M:

1. It is not easily understood.
2. It is difficult to calculate.
3. It is only an abstract figure and may not be the action of the series.

CHECK YOUR PROGRESS – 1

1. What the 3 measures of central tendency?
2. What is the formula for arithmetic mean under direct method?
3. Mention 2 merits of geometric mean
4. What do you mean by harmonic mean?

2.4 MEDIAN

The number of students in your classroom, the money your parents earns, the temperature in your city is all important numbers. But how can you get the information of the number of students in your school or the amount earned by the citizen of your entire city?

The median is that value of the variable which divides the

group into two equal parts, one part comprising all values greater and the other all values less than median.

Ungrouped data

Arrange the given values in the ascending or descending order.

If the number of value is odd, median is the middle value.

For example if we have the number of values 12, 15, 21, 27, 35. So the numbers are odd then taking the mean as the midpoint 21.

Median = $\frac{(n+1)^{th}}{2}$ term if n is odd

If the number of values is even, median is the mean of the middle two values.

For example if we have 12, 15, 21, 27, 35, 40. So the numbers are even then taking the mean of the numbers,

Median = Mean($\frac{(n)^{th}}{2}$ and $\frac{(n+1)^{th}}{2}$ terms)

So in the above example, take the mean of 21 and 27 and divide it by 2 which will give you 24.

Example:

The salaries of 8 employees who work for a small company are listed below. What is the median salary?

40,000; 29,000; 35,500; 31,000; 43,000; 30,000; 27,000; 32,000

Solution:

Arrange the data in ascending order

27,000; 29,000; 30,000; 31,000; 32,000; 35,500; 40,000; 43,000

Since there is an even number of items in the data set, we compute the median by taking the mean of the two middlemost numbers.

Mean ($\frac{(n)^{th}}{2}$ and $\frac{(n+1)^{th}}{2}$ terms) = $\frac{4^{th} + 5^{th}\ item}{2}$

= $\frac{31{,}000\ + 32{,}000}{2} = \frac{63{,}000}{2} = 31{,}500$

The median salary is 31,500

Example: 13

Find the median of the following set of points in a game: 15, 14, 10, 8, 12, 8, 16

Solution:

First arrange the values in an ascending order 8, 8, 10, 12, 14, 15, 16

The number of point values is 7, an odd number. Hence, the median is the value in the middle position.

$$\text{Median} = \left(\frac{n+1}{2}\right)^{th} \text{term}$$

$$= \left(\frac{7+1}{2}\right)^{th} \text{term} = 4^{th}$$

The median is 12

Grouped data:

In grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or continuous frequency distribution. Whatever may be the distribution, cumulative frequencies have to be calculated the total number of items.

Cumulative frequency: (cf)

Cumulative frequency of each class is the sum of the frequency of the class and the frequencies of the pervious classes, ie adding the frequencies successively, so that the last cumulative frequency gives the total number of items.

When the data follows a discrete set of values grouped by size, we use the formula $\frac{(n+1)^{th}}{2}$ item for finding the median. First we form a cumulative frequency distribution, and the median is that value which corresponds to the cumulative frequency in which $\frac{(n+1)^{th}}{2}$item lies.

Example: 14

The following frequency distribution is classified according to the number of students on different branches. Calculate the median number of leaves per branch.

No of Students	1	2	3	4	5	6	7
Number of Branches	2	11	15	20	25	18	10

Solution:

No of Students x	No of Branches f	Cumulative Frequency cf
1	2	2
2	11	13

3	15	28
4	20	48
5	25	73
6	18	91
7	10	101
Total	**101**	

Median = size of $\frac{(N+1)^{th}}{2}$item

= size of $\frac{(101+1)^{th}}{2}$ item

= 51^{th} item

Median = 5 because 51^{th} item corresponds to 5

Median for continuous grouped data

In case, the data is given in the form of a frequency table with class interval etc, then the following formula is used for calculating median in continuous grouped data

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

Where l = Lower limit of the median class

m = cumulative frequency preceding the median

c = width of the median class

f = frequency in the median class

N = total frequency

Example:

Calculate median from the following data

Class interval	0-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39
Frequency	5	8	10	12	7	6	3	2

Solution:

Class interval	Frequency f	True class interval	Cumulative frequency cf

0-4	5	0.5 - 4.5	5
5-9	8	4.5 - 9.5	13
10-14	10	9.5 - 14.5	23
15-19	12	14.5 - 19.5	35
20-24	7	19.5 - 24.5	42
25-29	6	24.5 - 29.5	48
30-34	3	29.5 - 34.5	51
35-39	2	34.5 - 39.5	53
	53		

$\frac{N}{2} = \frac{53}{2} = 26.5$

Here the cumulative frequency is greater than or equal to 26.5 is 14.5

$$\text{Median} = l + \frac{\frac{N}{2} - m}{f} \times c$$

l = 14.5

N/2 = 26.5

m = 23

f = 12

$$= 14.5 + \frac{(26.5 - 23)}{12} \times 5 = 14.5 + 1.46 = 15.96$$

Merits of Median:

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open end intervals.
3. Median can be located even if the data are incomplete.
4. Median can be located even for qualitative factors such as ability, honesty etc.

Demerits of Median:

1. A slight change in the series may bring drastic change in median value
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its

use in mean deviation.

4. It is not taken into account all the observation.

2.5 MODE

The mode is the most frequently occurring values or scores.The mode is useful when there are a lot of repeated values. There can be no mode, one mode, or multiple modes.

Its importance is very great in marketing studies where a manager is interested in knowing about the size, which has the highest concentration of items. For example, in placing an order foot shoes or ready-made garments the model size helps because the sizes and other sizes around in common demand.

Ungrouped Data:

For ungrouped values or a series of individual observation mode is often found by mere inspection

Example:

Find the mode for the following list of values: 13,18,13,14,13,16,14,21,13

Solution:

The mode is the number that is repeated more often than any other

Therefore the Mode = 13

In some cases the mode may be absent while in some cases there may be more than one mode.

Example:

Ms.Rossy asked students in her class how many siblings they each has.

Find the mode of the data : 0,0,0,1,1,1,1,2,2,2,2,3,3,4

Solution:

The modes are 1 and 2 siblings

Grouped Data

For Discrete distribution, the highest frequency and corresponding value of X is mode.

Continuous distribution:

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Where L is the lower class limit of the modal class

f_1 is the frequency of the modal class

f_0 is the frequency of the class preceding the modal class in the frequency table

f_2 is the frequency of the class succeeding the modal class in the frequency table

h is the class interval of the modal class

Example:18

Calculate mode for the following:

C-I	0-50	50-100	100-150	150-200	200-250	250-300	300-350	350-400	400 and above
f	5	14	40	91	450	87	60	38	15

Solution:

The highest frequency is 450 and corresponding class interval in 200 – 250, which is the modal class

Here L = 200, f_1 = 150, f_0=91, f_2=87, h=50

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

$$= 200 + \frac{150 - 91}{2 \times 150 - 91 - 87} \times 50$$

$$= \frac{2450}{122} = 200 + 24.18 = \mathbf{224.18}$$

Example: 19

Find the modal class and the actual mode of the data set below

Number	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30
Frequency	7	6	4	2	2	8	1	2	3	2

Solution:

Modal class = 10 – 12

$$\text{Mode} = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2}\right) \times h$$

Here L = 10, f_1 = 9, f_0 = 4, f_2 = 2, h = 3

$$= 10 + \frac{9 - 4}{2 \times 9 - 2 - 4} \times 3$$

$$= 10 + \frac{5}{12} \times 3 = 10 + 1.25 = 11.25$$

Mode = 11.25

Merits of mode:

1. It is easy to calculate and in some cases it can be located mere inspection.
2. Mode is not at all affected by extreme values
3. It can be calculated for open-end classes
4. It is usually an actual value of an important part of the series
5. In some circumstances it is the best representative of data

Demerits of mode:

1. It is not based on all observation
2. It is not capable of further mathematical treatment
3. Mode is ill defined generally it is not possible to find mode in some cases.
4. As compared with mean, mode is affected to a great extent by sampling fluctuations

It is unsuitable in cases where relative importance of items has to be considered.

2.6 PARTITION MEASURES

2.6.1 QUARTILES

The quartiles divide the distribution in four parts. There are three quartiles denoted by Q1, Q2 and Q3 divides the frequency distribution in to four equal parts

That is 25% of data will lie below Q1, 50% of data below Q2 and 75percent below Q3. Here Q2 is called the Median. Quartiles are obtained in almost the same way as median.

Ungrouped Data:

If the data set consist of n items and arranged in ascending order then

$$Q_1 = \left(\frac{n+1}{4}\right)^{th} \text{ item}, \quad Q_2 = \left(\frac{n+1}{2}\right)^{th} \text{ item and } Q_3 = 3\left(\frac{n+1}{4}\right)^{th} \text{ item}$$

Example:20

Compute quartiles for the data 25, 18, 30, 8, 15, 5, 10, 35, 40, 45.

Solution:

$$Q1 = \frac{(n+1)^{th}}{4}\text{item} = \frac{(10+1)^{th}}{4}\text{item} = (2.75)^{th} \text{ item}$$

$$= 2^{nd} \text{ item} + \frac{(3)^{rd}}{4}(3^{rd} \text{ item} - 2^{nd} \text{ item})$$

$= 8 + \frac{(3)}{4}(10 - 8) = 8 + 1.5$

Q1= 9.5

$Q3 = 3\frac{(n+1)^{th}}{4}$ item $= \frac{(10+1)^{th}}{4}$item $= 3 \times (2.75)^{th}$ item

$= (8.25)^{th}$ item

$= 2^{nd}$ item $+ \frac{(1)}{4}(9^{th}$ item $- 8^{th}$ item)

$= 35 + \frac{(1)}{4}(40 - 35) = 35 + 1.25$

Q3 = 36.25

Continuous series:

In the case of continuous series, find the cumulative frequency and then use the interpolation formula.

- Find Cumulative frequencies
- Find N / 4
- Q1 class is the class interval corresponding to the value of the cumulative frequency just greater than N / 4
- Q3 class is the class interval corresponding to the value of the cumulative frequency just greater than 3 N / 4

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1 \quad \text{and} \quad Q_3 = l_3 + \frac{3\left(\frac{N}{4}\right) - m_3}{f_3} \times C_3$$

Where N=Σf = total of all frequency values

l_1 = lower limit of the first quartile class

f_1 = frequency of the first quartile class

c_1 = width of the first quartile class

m_1 = cumulative frequency preceding the first quartile class

l_3 = lower limit of the 3rd quartile class

f_3 = frequency of the 3rd quartile class

m_3 = cumulative frequency preceding the 3rd quartile class

c_3 = width of the third quartile class

Example:

The marks secured by group of students in their internals.

Class	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
Frequency	4	3	2	1	5

Solution:

Class	Frequency f	Cumulative frequency cf
10 - 20	4	4
20 - 30	3	7
30 - 40	2	9
40 - 50	1	10
50 - 60	5	15

N / 4 = 15 / 4 = 3.75 which lies in 10 – 12

Lies in the group 10 – 20

$$Q_1 = l_1 + \frac{\frac{N}{4} - m_1}{f_1} \times c_1$$

$$= 10 + \frac{(3.75 - 0)}{4} \text{ x } 10 = 10 + 9.38 = \mathbf{19.38}$$

3N / 4 = 3x15 / 4 = 11.25 which lies in 50 -60

Therefore Q3 lies in the group 50 – 60

$$Q_3 = l_3 + \frac{\frac{3}{N} - m_3}{f_3} \times c_3$$

$$= 50 + \left(\frac{11.25 - 10}{5}\right) \text{ x } 10$$

= 50 + 2.5 = **52.5**

2.6.2 DECILES

These are the values which divide the total number of observation into 10 equal parts. They are D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9 and D_{10}.

Ungrouped Data:

Example:

Compute the D_7 for the data: 5, 24, 36, 12, 20, and 8.

Solution:

Arranging the given data in the ascending order 5,8,12,20,24,36

D5 = $\frac{(5(n+1))^{th}}{10}$ observation $= \frac{(5(6+1))^{th}}{10}$ observation = $(3.5)^{th}$ observation

= 3^{rd} item + ½ (4^{th} item - 3^{rd} item)

= 12 + ½ (20-12) = 12 + 4 **= 16**

Grouped Data:

Example:

Calculate the D_1 and D_7 for the given data

Class interval	0 -10	10-20	20-30	30-40	40-50	50-60	60-70
Frequency	5	7	12	16	10	8	4

Solution:

Class interval	Frequency f	Cumulative frequency cf
0 -10	5	5
10-20	7	12
20-30	12	24
30-40	16	40
40-50	10	50
50-60	8	58
60-70	4	62

$D_4 = (4N / 10)^{th}$ item $= (4 \times 62 / 10)^{th}$ item $= (24.8)^{th}$ item

This lies in the interval 30 – 40

$$D_4 = l + \frac{(4N / 10 - m)}{f} \times c$$

$$= 30 + \frac{(24.8-24)}{16} \times 10 \quad = \quad 30 + \frac{(0.8)}{16} \; (\underline{0.8}) \times 10$$

$= 30 + 0.5 \quad = \mathbf{30.5}$

$D_7 = (7N / 10)^{th}$ item

$= (7 \times 62 \; / 10)^{th}$ item

$= (43.4)^{th}$ item

This lies in the interval 40 – 50

$$D_4 = l + \frac{(7N / 10 - m)}{f} \times c$$

$$= 40 + \frac{(43.4-40)}{10} \times 10 \quad = 30 + \frac{(3.4)}{10} \times 10$$

$= 40 + 3.4 \quad = \mathbf{43.4}$

2.6.3 PERCENTILE

The percentile values divide the distribution into 100 parts each containing 1 percent of the cases. The percentile (P_k) is that value of the variable upto which lie exactly k% of the total number of observation

Relationship

$P_{25} = Q_1$

P_{50} = Median = Q_2

P_{75} = 3rd quartile = Q_3

Ungrouped Data:

Example: 24

The monthly income (in □ 1000) of 8 persons working in a factory. Find P_{30} income value 17, 21,14,36,10,25,15,29

Solution:

Arrange the data in the increasing order : 10, 14, 15, 17, 21, 25, 29, 36

n = 8

$$P_{30} = \left(\frac{30(n+1)}{100}\right)^{th} \text{ item}$$

$$= \left(\frac{30(8+1)}{100}\right)^{th} \text{ item}$$

$$= \left(\frac{30 \times 9}{100}\right)^{th} \text{ item} = 2.7^{th} \text{ item}$$

$= 2^{nd}$ item + 0.7(3^{rd} items - 2^{nd} items)

= 14 + 0.7 (15 -14)

= 14 + 0.7

P_{30} = 14.7

Grouped Data:

Example: 25

Find P_{53} for the following frequency distribution.

Class interval	**0-5**	**5-10**	**10-15**	**15-20**	**20-25**	**25-30**	**30-35**	**35-40**
Frequency	5	8	12	16	20	10	4	3

Solution:

Class interval	**Frequency**	**Cumulative frequency**
0-5	5	5

5-10	8	13
10-15	12	25
15-20	16	41
20-25	20	61
25-30	10	71
30-35	4	75
35 - 40	3	78
Total	**78**	

$$P_{53} = l + \left(\frac{53N/10 - m}{f}\right) \times c$$

$$= 20 + \left(\frac{41.34 - 41}{20}\right) \times 5 = 20 + 0.335 = \mathbf{20.335}$$

2.7 MEASURES OF DISPERSION

Dispersion is the extent till which a distribution can be stretched or squeezed. We can understand variation with the help of the following example:

Series I	Series II	Series III
10	2	10
10	8	12
10	20	8
$\sum X$ = 30	**30**	**30**

In all three series, the value of arithmetic mean is 10. On the basis of this average, we can say that the series are alike. If we carefully examine the composition of three series, we find the following differences:

(i) In case of 1st series, three items are equal; but in 2nd and 3rd series, the items are unequal and do not follow any specific order.

(ii) The magnitude of deviation, item-wise, is different for the 1st, 2nd and 3rd series. But all these deviations cannot be ascertained if the value of simple mean is taken into consideration.

(iii) In these three series, it is quite possible that the value of arithmetic mean is 10; but the value of median may differ from each other. This can be understood as follows;

Series I	Series II	Series III
10	2	8
10 median	8 median	10 median
10	20	12
$\sum X$ = 30	**30**	**30**

The value of Median' in 1st series is 10, in 2nd series = 8 and in 3rd series = 10. Therefore, the value of the Mean and Median are not identical.

(iv) As the average remains the same, the nature and extent of the distribution of the size of the items may vary. In other words, the structure of the frequency distributions may differ even though their means are identical.

2.7.1 PROPERTIES OF A GOOD MEASURE OF DISPERSION

There are certain pre-requisites for a good measure of dispersion:

1. It should be simple to understand.
2. It should be easy to compute.
3. It should be rigidly defined.
4. It should be based on each individual item of the distribution.
5. It should be capable of further algebraic treatment.

2.7.2 CHARACTERISTICS OF MEASURES OF DISPERSION

- A measure of dispersion should be rigidly defined
- It must be easy to calculate and understand
- Not affected much by the fluctuations of observations
- Based on all observations

2.7.3 CLASSIFICATION OF MEASURES OF DISPERSION

The measure of dispersion is categorized as:

(i) An absolute measure of dispersion:

It involves the units of measurements of the observations. For example, (i) the dispersion of salary of employees is expressed in rupees, and (ii) the variation of time required for workers is expressed in hours. Such measures are not suitable for comparing the variability of the two data sets which are expressed in different units of measurements

(ii) A relative measure of dispersion:

It is a pure number independent of the units of measurements. This measure is useful especially when the data sets are measured in different units of measurement

For example, a nutritionist would like to compare the obesity of school children in India and Africa. He collects data from some of the schools in these two countries. The weight is normally measured in kilograms in India and in pounds in Africa. It will be meaningless, if we compare the obesity of students using absolute measures. So it is sensible to compare them in relative measures.

2.8 RANGE

Raw Data:A range is the most common and easily understandable measure of dispersion. It is the difference between the largest and smallest observations in the data set

Range (R) = L - S

Grouped Data: The grouped frequency distribution of values in the data set, the range is the difference between the upper class limit of the last class interval and the lower class limit of the first class interval.

Coefficient of Range: The relative measure of range is called the coefficient of range

Coefficient of range = (L-S) / (L + S)

Example:

Find the value of range and its coefficient for the following data 49, 81, 36, 64, 121, 100.

Solution:

L = 121 : S = 36

Range : L – S = 121 – 36 = 85

Co-efficient of Range = (L-S) / (L+S) = 121-36 /121+36

= 85 / 157 = 0.5414

Example:

Calculate range and its coefficient from the following distribution.

x	10- 15	15 – 20	20 – 25	25 - 30
Frequency	4	10	16	8

Solution: L = 30, S = 10

Range = L - S = 30 – 10 = 20

Coefficient of Range = (L-S) / (L+S) = 30 - 10 / 30 + 10
= 20/ 40**= 0.5**

Merits of Range

- It is the simplest of the measure of dispersion
- Easy to calculate
- Easy to understand
- Independent of change of origin

Demerits of Range

- It is based on two extreme observations. Hence, get affected by fluctuations
- A range is not a reliable measure of dispersion
- Dependent on change of scale

2.9 QUARTILE DEVIATION

The quartiles divide a data set into quarters. The first quartile, (Q1) is the middle number between the smallest number and the median of the data. The second quartile, (Q2) is the median of the data set. The third quartile, (Q3) is the middle number between the median and the largest number. Quartile deviation is half of the difference between the first and third quartiles. Hence it is called as Semi Inter Quartile Range

Quartile deviation or semi-inter-quartile deviation is

$$Q = \frac{1}{2} \times (Q3 - Q1)$$

Coefficient of Quartile Deviation

Coefficient of Q.D = Q3 – Q1 / Q3 + Q1

Merits of Quartile Deviation

- All the drawbacks of Range are overcome by quartile deviation
- It uses half of the data
- Independent of change of origin
- The best measure of dispersion for open-end classification

Demerits of Quartile Deviation

- It ignores fifty percent of the data
- Dependent on change of scale
- Not a reliable measure of dispersion

Example:

Calculate the quartile deviation and its coefficient for the wheat production (in Kg) of 25 acres is given as : 1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750 and1885.

Solution: Arrange the observation in increasing order:

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

Q1 = value of (n+1) / 4 th item

= value of (20 +1) / 4 th item = value of (5.25)th item

= 5th item + 0.25 (6th item – 5th item)

= 1240 + 0.25 (1320 – 1240)

= 1240 + 20 = 1260

Q1 = 1260

Q3 = value of 3(n+1) / 4 th item

= value of 3(20 +1) / 4 th item = value of (15.75)th item

= 15th item + 0.75 (16th item –15th item)

= 1750 + 0.75 (1755 – 1750)

= 1750 + 3.75 = 1753.75

Q3 = 1753.75

Q.D = (Q3 – Q1) / 2 = (1753.75 – 1260) / 2 = 492.75 / 2

= 246.875

Coefficient of QD = (Q3 – Q1) / (Q3 + Q1)

= (1753.75 – 1260) / (1753.75 + 1260)

= 0.164

2.10 MEAN DEVIATION

The average deviation, it is defined as the sum of the deviations from an average divided by the number of items in a distribution The average can be mean, median or mode. Theoretically median is d best average of choice because sum of deviations from median is minimum, provided signs are ignored. However, practically speaking, arithmetic mean is the most commonly used average for calculating mean deviation and is denoted by the symbol MD.

Mean Deviation is of three types of series:

- Individual Data Series
- Discrete Data Series

- Continuous Data Series

Individual Data Series: For individual series, the Mean Deviation can be calculated using the following formula.

$$MD = \frac{1}{N}\sum|X - A| = \frac{\sum|D|}{N}$$

Where

MD = Mean deviation.

X = Variable values

A = Average of choices

N = Number of observations

Coefficient of Mean Deviation:

Mean deviation calculated by any measure of central tendency is an absolute measure. The purpose of comparing variation among different series, a relative mean deviation is require. The relative mean deviation are obtained by dividing the mean deviation by the average used for calculating mean deviation

The Coefficient of Mean Deviation can be calculated using

$$\textbf{Coefficient of MD} = \frac{\textbf{MD}}{\textbf{A}}$$

Example:

Calculate mean deviation and coefficient of mean deviation for the following individual data:

Items	28	72	90	140	210

Solution:

$$A = \frac{28 + 72 + 90 + 140 + 210}{5} = \frac{540}{5} = 108$$

Item X	Deviation |D|
28	80
72	36
90	18

140	32
210	102
	Σ|D|= 268

$$\text{Mean Deviation} = \text{MD} = \frac{1}{\text{N}}\sum |\text{X} - \text{A}| = \frac{\sum |D|}{N} = \frac{268}{5} = \mathbf{53.6}$$

$$\text{Coefficient of Mean Deviation} = \frac{\text{MD}}{\text{A}} = \frac{53.6}{108} = \mathbf{0.4963}$$

Discrete Data Series

For discrete series, the Mean Deviation can be calculated using

$$\boldsymbol{MD} = \frac{\sum \mathbf{f}\ |\mathbf{x} - \mathbf{Me}|}{\boldsymbol{N}} = \frac{\sum \mathbf{f}\ |\mathbf{D}|}{\boldsymbol{N}}$$

Where, N = Number of observations.

f = Different values of frequency f.

x = Different values of items.

Me = Median.

Coefficient of Mean Deviation

The Coefficient of Mean Deviation can be calculated using the following formula.

$$\textbf{Coefficient of MD} = \frac{\mathbf{MD}}{\mathbf{Me}}$$

Example: Calculate the mean deviation and for the following discrete data

Items	42	108	135	150	210
Frequency	6	15	3	3	9

Solution:

X_i	Frequency f_i	$f_i x_i$	$\|x_i - Me\|$	fi |xi – Me|
42	6	252	93	558
108	15	1620	27	405
135	3	405	0	0

150	3	550	15	45
210	9	1890	75	675
	N = 36			**$\Sigma f_i \|x_i - Me\|$** = 1683

$$\text{Median} = \frac{(\text{N} + 1)\text{th item}}{2} = \frac{(5+1)\text{th item}}{2} = \frac{6\text{th item}}{2} = 3\text{rd item}$$

$$= 135$$

$$\text{Mean Deviation} \quad == \frac{\sum f\,|x - Me|}{N} = \frac{\sum f\,|D|}{N} = \frac{1683}{36} = \mathbf{46.75}$$

$$\text{Coefficient of MD} = \frac{MD}{Me} = \frac{46.75}{135} = \mathbf{0.3463}$$

Continuous Data Series

The method of calculating mean deviation in a continuous series is same as the discrete series. In continuous series, find a midpoint of the various classes and take deviation of these points from the average selected

$$MD = \frac{\sum f\,|x - Me|}{N} = \frac{\sum f\,|D|}{N}$$

Where N = Number of observations.

f = Different values of frequency f.

x = Different values of items.

Me = Median.

Coefficient of Mean Deviation

The Coefficient of Mean Deviation can be calculated using the following formula.

$$\text{Coefficient of MD} = \frac{MD}{Me}$$

Example:

Find out the mean deviation from the given data

Age in years	**0-10**	**10-20**	**20-30**	**30-40**	**40-50**	**50-60**	**60-70**	**70-80**
No of persons	40	50	64	80	82	70	20	16

Solution:

Items	**Mid point xi**	**Frequency fi**	**fixi**	**|xi – Me|**	**fi |xi – Me|**	**Items**	**Mid point xi**	**Frequency fi**	**fixi**
0-10	5	40	200	31.47	1258.8	0-10	5	40	200
10-20	15	50	750	21.47	1073.5	10-20	15	50	750
20-30	25	64	1600	11.47	734.08	20-30	25	64	1600
30-40	35	80	2800	1.47	117.6	30-40	35	80	2800
40-50	45	82	3690	9.47	776.54	40-50	45	82	3690
50-60	55	70	3850	19.47	1362.9	50-60	55	70	3850
60-70	65	20	1300	29.47	589.4	60-70	65	20	1300
70-80	75	16	1200	39.47	631.52	70-80	75	16	1200
		N = 422	**$\Sigma f_i x_i$ =15390**						**$\Sigma f_i \|x_i – Me\|$ 6544.34**

$$\text{Median} = \frac{\Sigma \text{ fixi}}{\text{N}} = \frac{15390}{422} = \mathbf{36.47}$$

$$\text{Mean Deviation} == \frac{\sum f\,|x - \text{Me}|}{\text{N}} = \frac{\sum f\,|D|}{\text{N}} = \frac{6544.34}{422} = 15.5079$$

$$\text{Coefficient of MD} = \frac{\text{MD}}{\text{Me}} = \frac{15.5079}{36.47} = 0.4252$$

Merits of Mean Deviation:

- It is simple to understand and easy to compute.
- It is based on each and every item of the data.
- MD is less affected by the values of extreme items than the Standard deviation.

Demerits of Mean Deviation:

- The greatest drawback of this method is that algebraic signs are ignored while taking the deviations of the items.
- It is not capable of further algebraic treatments.
- It is much less popular as compared to standard deviation.

2.11 STANDARD DEVIATION

The concept of Standard Deviation was introduced by Karl Pearson in 1893. It is by far the most important and widely used measure of dispersion. Its significance lies in the fact that it is free from those defects which afflicted earlier methods and satisfies most of the properties of a good measure of dispersion. Standard Deviation is also known as root-mean square deviation as it is the square root of means of the squared deviations from the arithmetic mean.

The standard deviation is defined as the positive square root of the mean of the square deviations taken from the arithmetic mean of the data

Ungrouped data

x1 , x2 , x3 ... xn are the ungrouped data then standard deviation is calculated bythere are two methods of calculating standard deviation in an individual series

- Actual mean method
- Assumed mean method

Actual Mean Method:

$$\text{Standard deviation } \sigma = \frac{\sqrt{\Sigma(\text{X} - \bar{\text{X}})^2}}{\text{n}}$$

Example:

Calculate the standard deviation from the following data 28, 44, 18, 30, 40, 34, 24, 22.

Solution:

Deviations from actual mean

Values (X)	$X - \bar{X}$	$(X - \bar{X})^2$
28	-2	4
44	-14	196
18	-12	144
30	0	0
40	10	100
34	4	16
24	-6	36
22	-8	64
240		560

$$\bar{X} = \frac{240}{8} = 30$$

$$\sigma = \frac{\sqrt{\Sigma(X - \bar{X})^2}}{n} = \frac{\sqrt{560}}{8} = \sqrt{70} = \mathbf{8.3666}$$

Assumed Mean Method

This method is used when the arithmetic mean is fractional value. Taking deviations from fractional value would be a very difficult and tedious task. To save time and labour a short cut method is used; deviations are taken from a assumed mean.

$$Standard\ Deviation\ \sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

Example:

The marks obtained by the college students in statistics. Using the following data calculate standard deviation.

Students No:	1	2	3	4	5	6	7	8	9	10
Marks	53	58	46	67	32	70	35	68	88	99

Solution: Deviations from assumed mean

Students No	**Marks (x)**	**d = X – A (A=67)**	**d^2**
1	53	-14	196
2	58	-9	81
3	75	8	64
4	67	0	0
5	32	-35	1225
6	70	3	9
7	35	-32	1024
8	68	1	1
9	88	21	441
10	69	2	4
n = 10		**Σd = -55**	**Σd^2 = 3045**

$$\sigma = \sqrt{\frac{\Sigma d^2}{n} - \left(\frac{\Sigma d}{n}\right)^2}$$

$$= \sqrt{\frac{3045}{10} - \left(\frac{-55}{10}\right)^2} = \sqrt{304.5 - 30.25} = \sqrt{274.25}$$

$$= \mathbf{16.5605}$$

2.11.1 CALCULATION OF STANDARD DEVIATION

Discrete series: There are three methods for calculating standard deviation in discrete series. They are

a) Actual mean method
b) Assumed mean method
c) Step deviation method

Actual mean method

Calculate the mean of the series. Find the deviations for various items from the means and square the deviations and multiply by the respective frequency and total the product the formula to calculate actual mean method is

$$\sigma = \frac{\sqrt{\Sigma fd^2}}{\Sigma f}$$

If the actual mean is fractions, the calculation takes lot of time and labour; and as such this method is rarely used in practice

Assumed mean method

Here deviation is taken not from an actual mean but from an assumed mean. Also this method is used, if the given variable values are not in equal intervals.

$\sigma = \sqrt{\frac{\Sigma d^2}{f} - \left(\frac{\Sigma d}{f}\right)^2}$ where d = X – A, N = Σf

Example:

Calculate standard deviation from the following data:

X	20	22	25	31	35	40	42	45
f	5	12	15	20	25	14	10	6

Solution:

Deviation from assumed mean

x	**f**	**d = X-A (A=31)**	**d^2**	**fd**	**fd^2**
20	5	-11	121	-55	605
22	12	-9	81	-108	972
25	15	-6	36	-90	540
31	20	0	0	0	0
35	25	4	16	100	400
40	14	9	81	126	1134
42	10	11	121	110	1210

45	6	14	196	84	504
	N= 107			**Σfd = 167**	**Σ fd² = 5365**

Curriculum

NOTES

$$\sigma = \sqrt{\frac{\Sigma fd^2}{f} - \left(\frac{\Sigma fd}{f}\right)^2} = \sqrt{\frac{5365}{107} - \left(\frac{167}{107}\right)^2} = \sqrt{50.16 - 2.44} = \mathbf{6.91}$$

Step – deviation method:

If the variable values are in equal intervals, then we adopt this method

$$Standard\ Deviation\ \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}\ \text{X C}$$

Example:

The frequency distribution of marks in mathematics given in the table

Marks	30	40	50	60	70	80	90
No of students	8	12	20	10	7	3	2

Solution:

Marks x	**f**	**d= (x-50)/ 10**	**fd**	**fd²**
30	8	-2	-16	32
40	12	-1	-12	12
50	20	0	0	0
60	10	1	10	10
70	7	2	14	28
80	3	3	9	27
90	2	4	8	32
	N = 62		**Σfd = 13**	**Σfd² = 141**

$$Standard\ Deviation\ \sigma = \sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}\ X\ C$$

$$= \sqrt{\frac{141}{62} - \left(\frac{13}{62}\right)^2}\ X\ 10 = \ 1.4934\ X\ 10 = \ \mathbf{14.934}$$

Combined Mean and Combined Standard Deviation

Combined arithmetic mean can be computed if we know the mean and number ofitems in each group of the data.

$\overline{x}_1$, $\overline{x}_2$, σ_1, σ_2 are mean and standard deviation of two data sets having n_1 and n_2 asnumber of elements respectively.

$$\text{combined mean } \overline{x_{12}} = \frac{n_1\overline{x_1} + n_2\overline{x_2}}{n_1 + n_2} \quad \text{(if two data sets)}$$

$$\overline{x_{123}} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + n_3\overline{x_3}}{n_1 + n_2 + n_3} \quad \text{(if three data sets)}$$

Example:

Particulars regarding income of two company are given below:

	Company	
	A	B
No.of Employees	600	500
Average income	1500	1750
Standard deviation of income	10	9

Compute combined mean and combined standard deviation.

Solution:

Given $n_1 = 600$; $\overline{x}_1 = 1500$; $\sigma_1 = 10$

$n_2 = 500$; $\overline{x}_2 = 1750$; $\sigma_2 = 9$

$$= \frac{600 \text{ x } 1500\ + 500 \text{ x } 1750}{600 + 500} = \frac{900000\ + 875000}{1100}$$

$= \mathbf{1613.6363}$

Combined Standard Deviation:

$$\sigma_{12} = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}}$$

$d_1 = \bar{x}_{12} - \bar{x}_1 = 1613.6363 - 1500 = 113.6363$

$d_2 = \bar{x}_{12} - \bar{x}_2 = 1613.6363 – 1750 = -136.3637$

$$\sigma_{12} = \sqrt{\frac{600\ (100 + 12913.209) + 500\,(81 + 18595.0587)}{600 + 500}}$$

= 124.8488

Merits of Standard Deviation:

Among all measures of dispersion Standard Deviation is considered superior because it possesses almost all the requisite characteristics of a good measure of dispersion. It has the following merits:

- It is rigidly defined.
- It is based on all the observations of the series and hence it is representative.
- It is amenable to further algebraic treatment.
- It is least affected by fluctuations of sampling.

Demerits:

- It is more affected by extreme items.
- It cannot be exactly calculated for a distribution with open-ended classes.
- It is relatively difficult to calculate and understand.

2.12 COEFFICIENT OF VARIATION

The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean. The coefficient of variation represents the ratio of the standard deviation to the mean, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from one another.

Coefficient of Variation = (Standard Deviation / Mean) * 100.

$$CV = \left(\frac{\sigma}{\bar{x}}\right) X\,100$$

The coefficient of variation (CV) is a measure of relative variability. It is the ratio of the standard deviation to the mean (average). For example, the expression “The standard deviation is 15% of the mean is a CV.

The CV is particularly useful when you want to compare results from two different surveys or tests that have different measures or values. For example, if you are comparing the results from two tests that have

different scoring mechanisms. If sample A has a CV of 12% and sample B has a CV of 25%, you would say that sample B has more variation, relative to its mean.

Example:

Price of car in five years in two cities is given below:

Price in city A	Price in city B
20,00000	10,00000
22,00000	20,00000
19,00000	18,00000
23,00000	12,00000
16,00000	15,00000

Which city has more stable prices?

Solution:

City A			**City B**		
Price X (in lakhs)	**Deviation $\bar{x}$ = 20 dx**	**dx^2**	**Price Y (in lakhs)**	**Deviation $\bar{y}$ = 15 dx**	**dy^2**
20	0	0	10	-5	25
22	2	4	20	5	25
19	-1	1	18	3	9
23	3	9	12	-3	9
16	-4	16	15	0	0
$\Sigma x = 100$	**$\Sigma dx = 0$**	**$\Sigma dx^2 = 30$**	**$\Sigma y = 75$**	**$\Sigma dy = 0$**	**$\Sigma dy^2 = 68$**

City A: $\bar{x} = \Sigma x / n = 100 / 5 = 20$

$$\sigma x = \sqrt{\frac{\sum(X-\overline{X})^2}{n}} = \sqrt{\frac{\sum dx^2}{n}} = \sqrt{\frac{30}{5}} = 2.45$$

$$C.V.(X) = \left(\frac{\sigma}{\overline{x}}\right) X\,100 = \frac{2.45}{20}\,X\,100 = \mathbf{12.25}\%$$

City B: $\overline{x} = \Sigma x / n = 75 / 5 = 15$

$$\sigma y = \sqrt{\frac{\sum(y-\overline{y})^2}{n}} = \sqrt{\frac{\sum dy^2}{n}} = \sqrt{\frac{68}{5}} = 3.69$$

$$C.V.(Y) = \left(\frac{\sigma}{\overline{y}}\right) X\,100 = \frac{3.69}{15}\,X\,100 = \mathbf{24.6}\%$$

City A had more stable prices than City B, because the coefficient of variation is less in City A.

CHECK YOUR PROGRESS - 2

5. What is the coefficient of mean deviation?

6. What does standard deviation mean?

7. What is a measure of relative variability?

2.13 SUMMARY

- Measures of central tendency fail to reveal the degree of the spread out or the extent of the variability in individual items of the distribution. Dispersion is the extent till which a distribution can be stretched or squeezed
- A range is the most common and easily understandable measure of dispersion. It is the difference between the largest and smallest observations in the data set.

  Coefficient of range = (L-S) / (L + S)

- The quartiles divide a data set into quarters. The first quartile, (Q1) is the middle number between the smallest number and the median of the data. The third quartile, (Q3) is the middle number between the median and the largest number. Quartile deviation is half of the difference between the first and third quartiles. Hence it is called as Semi Inter Quartile Range
- The average deviation, it is defined as the sum of the deviations from an average divided by the number of items in a distribution The average can be mean, median or mode.
- The standard deviation is defined as the positive square root of the mean of the square deviations taken from the arithmetic mean of the data.

- The coefficient of variation (CV) is a statistical measure of the dispersion of data points in a data series around the mean.

2.14 KEY WORDS

Measures of dispersion, range quartile deviation, mean deviation, standard deviation, and coefficient of variable.

2.15 ANSWERS TO CHECK YOUR PROGRESS

1. Dispersion is the extent till which a distribution can be stretched or squeezed.
2. The grouped frequency distribution of values in the data set, the range is the difference between the upper class limit of the last class interval and the lower class limit of the first class interval.
3. Coefficient of Q.D = $Q_3 - Q_1$ / $Q_3 + Q_1$
4. All the drawbacks of Range are overcome by quartile deviation
5. Coefficient of MD = $\frac{MD}{Me}$
6. The standard deviation is defined as the positive square root of the mean of the square deviations taken from the arithmetic mean of the data.
7. Coefficient of variation (CV) $CV = \left(\frac{\sigma}{\bar{x}}\right) X\, 100$

2.16 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

- What is dispersion? How is it advantageous over the measures of central tendency?
- Write short notes on range
- What is Coefficient of variation? Explain
- Write about mean deviation

LONG ANSWER QUESTIONS

- Calculate mean deviation under assumed mean method

Marks	30	40	50	60	70	80	90
No of students	16	24	40	20	14	6	4

- Give a detained account on standard deviation.
- Calculate the quartile deviation and its coefficient for the corn production (in Kg) of 25 acres is given as: 1100, 1340, 1370, 1050, 1780, 1200, 2440, 1390, 1480, 1780, 1783, 1542, 1970, 1680, 1775, 1320, 1680, 1770, 1780 and1889.

2.17 FURTHER READINGS

1. Levin, Richard I. and David S. Rubin: Statistics for Management, PrenticeHall, New Delhi.
2. Watsman terry J. and Keith Parramor: Quantitative Methods in FinanceInternational, Thompson Business Press, London.
3. Hooda, R. P.: Statistics for Business and Economics, Macmillan, New Delhi.
4. Hein, L. W. Quantitative Approach to Managerial Decisions, Prentice Hall, NJ

UNIT III - PROBABILITY

Structure

3.0 INTRODUCTION

In our day to day life the "probability" or "chance" is very commonly used term. Sometimes, we use to say "Probably it may rain tomorrow", "Probably Mr. X may come for taking his class today", "Probably you are right". All these terms, possibility and probability convey the same meaning. But in statistics probability has certain special connotation unlike in Layman's view.

The theory of probability has been developed in 17th century. It has got its origin from games, tossing coins, throwing a dice, drawing a card from a pack. In 1954 Antoine Gornband had taken an initiation and an interest for this area.

After him many authors in statistics had tried to remodel the idea given by the former. The "probability" has become one of the basic tools of statistics. Sometimes statistical analysis becomes paralyzed without the theorem of probability. "Probability of a given event is defined as the expected frequency of occurrence of the event among events of a like sort." (Garrett)

The probability theory provides a means of getting an idea of the likelihood of occurrence of different events resulting from a random experiment in terms of quantitative measures ranging between zero and

one. The probability is zero for an impossible event and one for an event which is certain to occur.

3.1 OBJECTIVES

The students will be able to understand

- The important terms in probability
- Concept of conditional probability, addition theorem and multiplication theorem.
- Baye's theorem and its applications

3.2 IMPORTANT TERMS

1. **Probability or Chance:** Probability or chance is a common term used in day-to-day life. For example, we generally say, 'it may rain today'. This statement has a certain uncertainty. Probability is quantitative measure of the chance of occurrence of a particular event.
2. **Experiment:** An experiment is an operation which can produce well-defined outcomes**.**

3. **Random Experiment:** If all the possible outcomes of an experiment are known but the exact output cannot be predicted in advance, that experiment is called a random experiment**.**
   Examples: Tossing of a fair coin**:** When we toss a coin, the outcome will be either Head (H) or Tail (T)

4. **Trial :** Any particular performance of a random experiment is called trial
   Example: Tossing 4 coins, rolling a die, picking ball from a bag containing 10 balls of which 4 is red and 6 is blue.

5. **Event :**Any subset of a Sample Space is an event. Events are generally denoted by capital letters A, B , C, D etc.
   Examples:
   i. When a coin is tossed, outcome of getting head or tail is an event

   Types of Events:

   - **Simple Events:** In the case of simple events, we take the probability of occurrence of single events**.**

   Examples: Probability of getting a Head (H) when a coin is tossed

   - **Compound Events:** In the case of compound events, we take the probability of joint occurrence of two or more events
     Examples: When two coins are tossed, probability of getting a Head (H) in the first toss and getting a Tail

(T) in the second toss..

6. **Sample Space :**Sample Space is the set of all possible outcomes of an experiment. It is denoted by S**.**
   Examples : When a coin is tossed, S = {H, T} where H = Head and T = Tail
7. **Mutually Exclusive Events:** Two or more than two events are said to be mutually exclusive if the occurrence of one of the events excludes the occurrence of the other
   Example : When a coin is tossed, we get either Head or Tail. Head and Tail cannot come simultaneously. Hence occurrence of Head and Tail are mutually exclusive events.

8. **Equally Likely Events**: Events are said to be equally likely if there is no preference for a particular event over the other.

   Examples: When a coin is tossed, Head (H) or Tail is equally likely to occur.
9. **Independent Events:** Events can be said to be independent if the occurrence or non-occurrence of one event does not influence the occurrence or non-occurrence of the other.

   Example:

   i. When a coin is tossed twice, the event of getting Tail(T) in the first toss and the event of getting Tail(T) in the second toss are independent events. This is because the occurrence of getting Tail(T) in any toss does not influence the occurrence of getting Tail(T) in the other toss.
10. **Exhaustive Events:** Exhaustive Event is the total number of all possible outcomes of an experiment.

    Examples: When a coin is tossed, we get either Head or Tail. Hence there are 2 exhaustive events.
11. **Favorable Events:** The outcomes which make necessary the happening of an event in a trial are called favorable events.
    Examples:if two dice are thrown, the number of favorable events of getting a sum 5 is four, i.e., (1, 4), (2, 3), (3, 2) and (4, 1).

3.3 TYPES OF PROBABILITY

1. **Classical Approach (Priori Probability):**
   According to this approach, the probability is the ratio of favorable events to the total no. of equally likely events. In tossing a coin the probability of the coin coming down ids 1, of the head coming up is ½ and of the tail coming up is ½.
   The probability of one event as ‘P’ (success) and of the other event as ‘q’ (failure) as there is no third event.

$p = \frac{\text{Number of favourable cases}}{\text{Total number of equally likely cases}}$

If an event can occur in ‘a’ ways and fail to occur in ‘b’ ways and these are equally to occur, then the probability of the event occurring, a/a+b is
denoted by p. Such probabilities are known as unitary or theoretical or mathematical probability.
p is the probability of the event happening and q is the probability of its not happening.

$p = \frac{a}{a+b}$ and $q = \frac{b}{a+b}$

Hence $p+q = \frac{a+b}{a+b}$

Therefore p+q = 1

Probabilities can be expressed either as ratio,fraction or percentage, such as ½ or 0.5 or 50%.**Example:** Tossing of a coin.

Limitations:

- This definition is confined to the problemsof games of chance only and can notexplain the problem other than the gamesof chance.
- This method can not be applied, when theoutcomes of a random experiment are notequally likely.
- The classical definition is applicable onlywhen the events are mutually exclusive.

2. **Relative Frequency Theory of Probability:**

In this approach, the probability of happening ofan event is determined on the basis of past experience or on the basis of relative frequency of success in the past.

Example: If a machine produces 100 articles in the past, 2 articles were found to be defective, and then the probability of the defective articles is 2/100 or 2%.
The relative frequency obtained on the basis of past experience can be shown to come to very close to the classical probability.

Limitations:

- The experimental conditions may not remain essentially homogeneous and identical in a large number of

repetitions of the experiment.

- o The relative frequency m/n, may not attain aunique value no matter however large.
- o Probability p(A) defined can never be obtainedin practice. We can only attempt at a closeestimate of p(A) by making N sufficiently large.

3. **Subjective Approach :**

The subjective approach is also known as subjective theory of probability. The probability of an event is considered as a measure of one's confidence in the occurrence of that particular event

This theory is commonly used in business decision making. The decision reflects the personality of the decision maker. Persons may arrive at different probability assignment because of differences in value at experience etc. The personality of the decision maker is reflected in a final decision. The decision under this theory is taken on the basis of the available data plus the effects of other factors many of which may be subjective in nature.

Example:A student would top in B. Com Exam this year.

A subjective would assign a weight between zero and one to this event according to his belief for its possible occurrence.

4. **Axiomatic Approach:**

The probability calculations are based on the axioms. The axiomatic probability includes the concept of both classical and empirical definitions of probability.

The approach assumes finite sample spaces and is based on the following three axioms:

i) The probability of an event ranges from 0 to 1.If the event cannot take place its probability shall be '0' and if it is bound to occur its probability is'1'.

ii) The probability of the entire sample space is 1, i.e. p(S)=1.

iii) If A and B are mutually exclusive events then the probability of occurrence of either A or B denoted byp(AUB) = p(A) + p(B)

iv) If A and B are happening together events then the probability of occurrence of probability of A intersection B denoted by p (A∩ B) = p(A) . p(B)

3.4 BASIC RELATIONSHIPS OF PROBABILITY

There are some basic probability relationships that can be used to compute the probability of an event without knowledge of all the sample point probabilities.

Complement of an Event:The complement of any even A is the even (not A),i.e, the event that A does not occur. The event A and its complement (not A) are mutually exclusive and exhaustive.it is denoted A′ , A^c or $\bar{A}$

Union of Two Events: the union of events A and B is the event containing all sample points that are in A or B or both. It is denoted by AUB

Intersection of Two Events: The intersection of events A and B is the set of all sample points that are in both A and B. it is denoted by A∩ B

○ **Mutually Exclusive Events**: two sets are mutually exclusive (also called disjoint) if they do not have any elements in common; they need not together comprise the universal set.

3.5 ADDITION THEOREM OF PROBABILITY

The probability of an event in a random experiment as well as axiomatic approach formulated by Russian Mathematician A.N. Kolmogorov and observed that probability as a function of outcomes of an experiment. By now you know that the probability P(A) of an event A associated with a discrete sample space is the sum of the probabilities assigned to the sample points in A as discussed in axiomatic approach of probability. Here we will learn Addition Theorem of Probability to find probability of occurrence for simultaneous trials under two conditions when events are mutually exclusive and when they are not mutually exclusive.

1. Addition Theorem For Mutually Exclusive Events

Statement: If A and B are two mutually exclusive events, then the probability of occurrence of either A or B is the sum of the individual probabilities of A and B. Symbolically

$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B)$$

Proof : Let N be the total number exhaustive and equally likely cases of an experiment. Let m_1 and m_2 be the number of cases favourable to the happening of events A and B respectively. Then

$$P(A) = \frac{n(A)}{n(S)} = \frac{m_1}{N}$$

and

$$P(B) = \frac{n(B)}{n(S)} = \frac{m_2}{N}.$$

Since the events A and B are mutually exclusive, the total number of events favorable to either A or B i.e. $n(A \cup B) = m_1 + m_2$, then

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N} = \frac{m_1 + m_2}{N} = \frac{m_1}{N} + \frac{m_2}{N} = P(A) + P(B)$$

Example 1: A card is drawn at random from a pack of 52 cards. Find the probability that the drawn card is either a club or an ace of diamond.

Solution : Let A : Event of drawing a card of club and

B: Event of drawing an ace of diamond

The probability of drawing a card of club $P(A) = \frac{13}{52}$

The probability of drawing an ace of diamond $P(B) = \frac{1}{52}$

Since the events are mutually exclusive, the probability of the drawn card being a club or an ace of diamond is:

$$P(A \cup B) = P(A) + P(B) = \frac{13}{52} + \frac{1}{52} = \frac{14}{52} = \frac{7}{26}$$

2. Addition Theorem For Non-Mutually Exclusive Events

The addition theorem discussed above is not applicable when the events are not mutually exclusive. For example, if one card is drawn at random from a pack of 52 cards then in order to find the probability of either a spade or a king card, it cannot be calculated by simply adding the probabilities of spade and king card because the events are not mutually exclusive as there is one card which is a spade as well as a king. Thus, the events are not mutually exclusive; therefore, the addition theorem is

modified as:

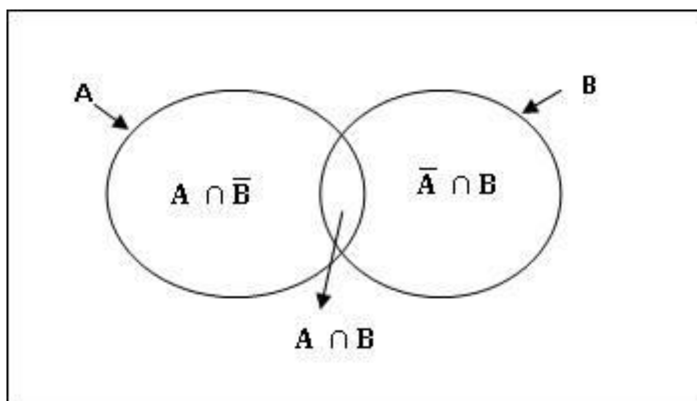
Statement: If A and B are not mutually exclusive events, the probability of the occurrence of either A or B or both is equal to the probability that event A occurs, plus the probability that event B occurs minus the probability of occurrence of the events common to both A and B. In other words the probability of occurrence of at least one of them is given by

$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B) - P(A \cap B)$$

Proof: Let us suppose that a random experiment results in a sample space S with N sample points (exhaustive number of cases). Then by definition

$$P(A \cup B) = \frac{n(A \cup B)}{n(S)} = \frac{n(A \cup B)}{N}$$

Where n(A∪B) is the number of occurrences (sample points) favorable to the event (A∪B)



Addition theorem for non-mutually exclusive events

From the above diagram, we get:

$$P(A \cup B) = \frac{[n(A) - n(A \cap B)] + n(A \cap B) + [n(B) - n(A \cap B)]}{N}$$

$$= \frac{n(A) + n(B) - n(A \cap B)}{N}$$

$$= \frac{n(A)}{N} + \frac{n(B)}{N} - \frac{n(A \cap B)}{N}$$

$$= P(A) + P(B) - P(A \cap B)$$

Example 2

A card is drawn at random from a pack of 52 cards. Find the probability

that the drawn card is either a spade or a king.

Solution: Let A: Event of drawing a card of spade and

B: Event of drawing a king card

The probability of drawing a card of spade

$$P(A) = \frac{13}{52}$$

The probability of drawing a king card

$$P(B) = \frac{4}{52}$$

Because one of the kings is a spade card also therefore, these events are not mutually exclusive. The probability of drawing a king of spade is

$$P(A \cap B) = \frac{1}{52}$$

So, the probability of the drawing a spade or king card is:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

3.6 MULTIPLICATION THEOREM OF PROBABILITY

In the addition theorem of probability for mutually exclusive events as well as for those events which are not mutually exclusive. In many situations we want to find the probability of simultaneous occurrence of two or more events. Sometimes the information is available that an event A has occurred and one is required to find the probability of occurrence of another event B utilizing the information about event A. Such a probability is known as conditional probability. Here we shall discuss the important concept of conditional probability of an event which will be helpful in understanding the concept of multiplication theorem of probability as well as independence of events.

1. **Multiplication Theorem for Independent Events**

Statement: This theorem states that if two events A and B are independent then the probability that both of them will occur is equal to the product of their individual probabilities.

P (AB) = P (A∩B) = A (A and B) = P (A). P (B)

Proof

If an event A can happen in n_1 ways out of which a_1 are favorable and the

event B can happen in n_2 ways out of which a_2 are favorable, we can combine each favorable event in the first with each favorable event in the second case. Thus, the total number of favorable cases is a_1 x a_2. Similarly, the total number of possible cases is n_1 x n_2. Then by definition the probability of happening of both the independent events is

$$P(A \cap B) = P(A \text{ and } B) = \frac{a_1 X a_2}{n_1 X n_2} = \frac{a_1}{n_1} X \frac{a_2}{n_2} = P(A) X P(B)$$

$$\text{as } P(A) = \frac{a_1}{n_1} \ \& \ P(B) = \frac{a_2}{n_2}$$

Similarly we can extend the theorem to three events

$P(A \cap B \cap C) = P(A) . P(B/A) . P(C/A . B)$

Example 1. From a pack of 52 cards, two cards are drawn at random one after the other with replacement. What is the probability that both cards are kings?

Solution:

The probability of drawing a king P (A) $=\frac{4}{52}$

The probability of drawing again the king after replacement P (B) = $\frac{4}{52}$

Since the two events are independent, the probability of drawing two kings is:

$$P(A \text{ and } B) = P(A \cap B) = P(A) \times P(B) = \frac{4}{52} \times \frac{4}{52} = \frac{1}{169}$$

2. **Multiplication Theorem of Probability for Dependent Events**

Statement: The probability of simultaneous happening of two events A and B is given by:

$$P(A \cap B) = P(A).P(B|A);\ P(A) \neq 0$$

$$P(B \cap A) = P(B).P(A|B);\ P(B) \neq 0$$

Where P (B|A) is the conditional probability of happening of B under the condition that A has happened and P (A|B) is the conditional probability of happening of A under the condition that B has happened.

Proof:

Let A and B be the events associated with the sample space S of a random experiment with exhaustive number of outcomes (sample points) N, i.e., n(S) = N. Then by definition

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)}$$

For the conditional event A|B (i.e., the happening of A under the condition that B has happened), the favorable outcomes (sample points) must be out of the sample points of B. In other words, for the event A|B, the sample space is B and hence

$$P(A|B) = \frac{n(A \cap B)}{n(B)}$$

Similarly, we have

$$P(B|A) = \frac{n(B \cap A)}{n(A)}$$

On multiplying and dividing equation (1) by n (A), we get

$$P(A \cap B) = \frac{n(A)}{n(S)} \times \frac{n(A \cap B)}{n(A)}$$

$$= P(A).P(B|A)$$

Also

$$P(A \cap B) = \frac{n(B)}{n(S)} \times \frac{n(A \cap B)}{n(B)}$$

$$= P(B).P(A|B)$$

Example

A bag contains 5 white and 8 red balls. Two successive drawings of 3 balls are made such that (a) the balls are replaced before the second drawing, and (b) the balls are not replaced before the second draw. Find the probability that the first drawing will give 3 white and the second 3 red balls in each case.

Solution:

(a) When balls are replaced.

Total balls in the bag = 8 + 5 = 13

3 balls can be drawn out of total of 13 balls in $^{13}C_3$ ways.

3 white balls can be drawn out of 5 white balls in $^{5}C_3$ ways.

Probability of 3 white balls = $P(3W) = \frac{^{5}C_3}{^{13}C_3} = \frac{10}{286}$

Since the balls are replaced after the first draw so again there are 13 balls in the bag 3 red balls can be drawn out of 8 red balls in $^{8}C_3$ ways.

Probability of 3 red balls = $P(3R) = \frac{^8C_3}{^{13}C_3} = \frac{56}{286}$

Since the events are independent, the required probability is:

$$P(3W \text{ and } 3R) = P(3W) \times P(3R) = \frac{^5C_3}{^{13}C_3} \times \frac{^8C_3}{^{13}C_3} = \frac{10}{286} \times \frac{56}{286} = \frac{140}{20{,}449}$$

(b) When the balls are not replaced before second draw

Total balls in the bag = 8 + 5 = 13

3 balls can be drawn out of 13 balls in $^{13}C_3$ ways.

3 white balls can be drawn out of 5 white balls in $^5C_3$ ways.

The probability of drawing 3 white balls = $P(3W) = \frac{^5C_3}{^{13}C_3}$

After the first draw, balls left are 10, 3 balls can be drawn out of 10 balls in $^{10}C_3$ ways.

3 red balls can be drawn out of 8 balls in $^8C_3$ ways. Probability of drawing 3 red balls = $\frac{^8C_3}{^{10}C_3}$.

Since both the events are dependent, the required probability is:

$$P(3W \text{ and } 3R) = P(3W) x P(3R|3W) = \frac{^5C_3}{^{13}C_3} \times \frac{^8C_3}{^{10}C_3} = \frac{5}{143} \times \frac{7}{15} = \frac{7}{429}$$

3.7 CONDITIONAL PROBABILITY

When the occurrence of an event A and are required to find out the probability of the occurrence of another event B. Two events A and B are said to be dependent when event A can occur only when event B is known to have occurred (or vice versa). The probability attached to such an event is called the conditional probability and is denoted by P (A|B) or in other words, probability of A given that B has occurred. For example, if we want to find the probability of an ace of spade if we know that card drawn from a pack of cards is black. Let us consider another problem relating to dairy plant. There are two lots of full cream packets A and B, each containing some defective packets. A coin is tossed and if it turns up with its head upside lot A is selected and if it turns with tail up, lot B is selected. In this problem we are interested to know the probability of the event that a milk packet selected from the lot obtained in this manner is defective.

Definition: If two events A and B are dependent, then the conditional probability of B given that event A has occurred is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}, \text{if } P(A) \geq 0$$

Let us consider the experiment of throwing of a die once. The sample space of this experiment is {1, 2, 3, 4, 5, and 6}.

Let E_1: an even number and E_2: multiple of 3 p.

Then E_1: {2, 4, 6} and E_2: {3, 6}.

Hence, P (E_1) = 3/6 = 1/2 and P (E_1) = 2/6 =1/3

In order to find the probability of occurrence of E_2 when it is given that E_1 has occurred We know that in a single throw of die2or4or 6has come up. Out of these only 6 is favorable to E_2. So the probability of occurrence of E_2 when it is given that E_1 has occurred is equal to 1/3. This probability of E_2 when E_1 has occurred is written as

P ($E_2|E_1$). Here we find that P ($E_2|E_1$) =P (E_2).

Let us consider the event

E_3: a number greater than 3 then E_3:{4,5,6} and P(E_3)=3/6=1/2 Out of 2,4 and 6, two numbers namely 4 and 6 are favorable to E_3.

Therefore, P ($E_3|E_1$) =2/3.

The events of the type E_1 and E_2 are called independent events as the occurrence or non-occurrence of E_1 does not affect the probability of occurrence or non-occurrence of E_2. The events E_1 and E_3 are not independent.

3.7.1 Combined Use Of Addition And Multiplication Theorem

In probability both addition and multiplication theorems are used simultaneously. The following examples illustrate the combined use of addition and multiplication theorems.

Example

A bag contains 5 white and 4 black balls. A ball is drawn from this bag and is replaced and then second draw of a ball is made. What is the probability that two balls are of different colors.

Solution: There are two possibilities

i) First ball is white and the second ball drawn is black.

ii) First ball is black and the second ball drawn is white.

Since the events are independent, so by using multiplication theorem we have

i) Probability of drawing First ball white and the second ball

black = $\frac{5}{9} \times \frac{4}{9} = \frac{20}{81}$

ii) Probability of drawing First ball black and the second ball white $= \frac{4}{9} \times \frac{5}{9} = \frac{20}{81}$

Since these probabilities are mutually exclusive, by using addition theorem

Probability that two balls are of different colors = $\frac{20}{81} \times \frac{20}{81} = \frac{40}{81}$

CHECK YOUR PROGRESS

1. What is sample space?
2. What is an event?
3. Write the formula for addition probability theorem
4. Mention the types of probability
5. How Baye's theorem is calculated

3.8 BAYES' THEOREM AND ITS APPLICATIONS

There are many situations where the ultimate outcome of an experiment tdepends on what happens in various intermediate stages. This issue is resolved by the Bayes'

There is a very big difference between **P(A | B) and P(B | A)**

Suppose that a new test is developed to identify people who are liable to suffer from some genetic disease in later life. Of course, no test is perfect; there will be some carriers of the defective gene who test negative, and some non-carriers who test positive. So, for example, let A be the event 'the patient is a carrier', and B the event 'the test result is positive'.

The scientists who develop the test are concerned with the probabilities that the test result is wrong, that is, with P(B | A′) and P(B′ | A). However, a patient who has taken the test has different concerns.

If I tested positive, what is the chance that I have the disease?

If I tested negative, how sure can I be that I am not a carrier? In other words, **P(A | B) and P(A ′ | B ′**).

These conditional probabilities are related by Bayes' Theorem:

Let A and B be events with non-zero probability. Then

$$P(A \mid B) = \frac{P(B \mid A)\cdot P(A)}{P(B)}$$

The proof is not hard. We have

P(A | B)·P(B) = P(A∩B) = P(B | A)·P(A),
using the definition of conditional probability twice. (Note that we need both A and B to have non-zero probability here.) Now divide this equation by P(B) to get the result.

If P(A) ≠ 0,1 and P(B)≠ 0, then

$$\mathbf{P(A \mid B)} = \frac{\mathbf{P(B \mid A) \cdot P(A)}}{\mathbf{P(B \mid A) \cdot P(A) + P(B \mid A') \cdot P(A')}}.$$

Bayes' Theorem is often stated in this form.

Example

Consider the clinical test described at the start of this section. Suppose that 1 in 1000 of the population is a carrier of the disease. Suppose also that the probability that a carrier tests negative is 1%, while the probability that a non carrier tests positive is 5%. (A test achieving these values would be regarded as very successful.) Let A be the event 'the patient is a carrier', and B the event 'the test result is positive'. We are given that P(A) = 0.001 (so that P(A′) = 0.999), and that
P(B | A) = 0.99, P(B | A ′) = 0.05.

(a) A patient has just had a positive test result. What is the probability that the patient is a carrier? The answer is

$$P(A \mid B) = \frac{\mathbf{P(B \mid A) \cdot P(A)}}{\mathbf{P(B \mid A) \cdot P(A) + P(B \mid A') \cdot P(A')}}$$

$$= \frac{0.99 \times 0.001}{(0.99 \times 0.001) + (0.05 \times 0.999)}$$

$$= \frac{0.00099}{0.05094} = 0.0194.$$

(b) A patient has just had a negative test result. What is the probability that the patient is a carrier? The answer is

$$P(A \mid B') = \frac{P(B' \mid A)P(A)}{P(B' \mid A)P(A) + P(B' \mid A')P(A')}$$

$$= \frac{0.01 \times 0.001}{(0.01 \times 0.001) + (0.95 \times 0.999)}$$

$$= \frac{0.00001}{0.94095} = 0.00001.$$

3.9 SUMMARY

- Bayes' Theorem is often stated in the form. If P(A) ≠ 0,1 and P(B)≠ 0, then

$$\mathbf{P(A \mid B)} = \frac{\mathbf{P(B \mid A) \cdot P(A)}}{\mathbf{P(B \mid A) \cdot P(A) + P(B \mid A') \cdot P(A')}}$$

- **Conditional Probability:** Two events A and B are said to be dependent when event A can occur only when event B is known to have occurred (or vice versa).
- **Multiplication Probability** :The probability of simultaneous occurrence of two or more events
- **Addition Probability**: If A and B are not mutually exclusive events, the probability of the occurrence of either A or B or both is equal to the probability that event A occurs, plus the probability that event B occurs minus the probability of occurrence of the events common to both A and B
- **Types Of Probability:** Axiomatic Approach,Classical Approach ,Relative Frequency Theory of Probability,Subjective Approach

3.10 KEY WORDS

Probability, Sample, Events, Variables, Addition theorem, Multiplication theorem, Axiomatic approach, Classical approach, Relative frequency theory, Subjective approach, Baye's theorem

3.11 ANSWER TO CHECK YOUR PROGRESS

1. **Sample Space :**Sample Space is the set of all possible outcomes of an experiment. It is denoted by S
2. **Event :**Any subset of a Sample Space is an event. Events are generally denoted by capital letters A, B , C, D etc.
3. Addition Theorem For Mutually Exclusive Events
   $\mathbf{P(A \cup B) = P(A\ or\ B) = P(A) + P(B)}$
   Addition Theorem For Non-Mutually Exclusive Events
   $\mathbf{P(A \cup B) = P(A\ or\ B) = P(A) + P(B) - P(A \cap B)}$
4. **Types Of Probability:** Axiomatic Approach,Classical Approach ,Relative Frequency Theory of Probability,Subjective Approach
5. Bayes' Theorem is often stated in the form. If P(A) ≠ 0,1 and P(B)≠ 0, then

$$\mathbf{P(A \mid B) = \frac{P(B \mid A)\cdot P(A)}{P(B \mid A)\cdot P(A) + P(B \mid A')\cdot P(A')}}$$

3.12 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTION:

1. Define probability
2. What are sample space
3. Define random variable
4. State the Baye's theorem
5. Explain mutually exclusive event

LONG ANSWER QUESTIONS:

1. Define probability and bring out the importance of probability
2. Distinguish between independent and dependents events
3. Explain briefly Baye's theorem

4. If 20% of the bottles produced by machine are defective, determine the probability that out of 4 bottles (i) 0, (ii) 1, (iii) at most 2 bottles will be defective

3.13 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers andDistributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing CompanyLtd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw HillPublishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., NewDelhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons.,NewDelhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.

UNIT IV - PROBABILITY DISTRIBUTION

Structure

4.0 INTRODUCTION

A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence. It describes the range of possible values that a random variable can attain and the probability that the value of the random variable is with any subset of that range. For example if X is a random variable then denote by P(X) to be the probability that X Occurs. It must be the case that 0<= P(X) <= 1 for each value of X and ΣP(X) = 1 (the sum of all the probabilities is 1)

4.1 OBJECTIVES

The students will be able to understand

- The random variable and its types in probability distribution
- Concept of Binomial Distribution, Poisson Distribution and Normal Distribution
- Concept of Mean and Standard deviation of Binomial and Poisson Distribution

4.2 Random variable

A random variable is defined as a real number X connected with the outcome of a random experiment E. For example, if E consists of three tosses of a coin, we may consider random variable X which denotes the number of heads (0, 1, 2 or 3)

Outcome :	HHH	HTH	THH	THH	HTT	THT	TTH	TTT

Value of X :	3	2	2	2	1	1	1	0

Thus, to every outcome there corresponds a real number X(w). Since the points of the sample space corresponds to outcomes, this means that a real number, which we denote by X(w), is defined for each w∈S and let us denote them by w_1, w_2,….,w_8 i.e.X(w_1) = 3, X(w_2)=2,…., X (w_8) =0. Thus,, we define a random variable as a real valued function whose domain is the sample space associated with a random experiment and range is the real line. Generally it is denoted by capital letters X,Y, Z, --- etc.

4.3 TYPES OF RANDOM VARIABLE

1. Discrete random variable

If a random variable X assumes only a finite or countable set of values, it is called a discrete random variable. In other words, a real valued function defined on a discrete sample space is called a discrete random variable. In case of discrete random variable we usually talk of values at a point. Generally it represents counted data. For example, number of defective milk packet in a milk plant, number of students in a class etc.

2. Continuous random variable

A random variable is said to be continuous if it can assume infinite and uncountable set of values. A continuous random variable is in which different values cannot be put in one to one correspondence with a set of positive integers. For example, weight of baby elephant take any possible value in the interval of 160 kg to 260 kg, say 189 kg or 189.4356 kg; likewise, marks scored by the students in a class etc. In case of continuous random variable we usually take the values in a particular interval. Continuous random variables represent measured data.

Probability Distribution of a Random Variable

The concept of probability distribution is equivalent to the frequency distribution. It depicts how total probability of one is distributed among various values which a random variable can take.

Mean and Variance of a Random variable

Let X denotes the random variable which assumes values x_1,x_2,---,x_n with corresponding probabilities p_1,p_2,---,p_n .Then the probability distribution be as follow:

X:	x_1	x_2	... ...	x_n
P(X):	p_1	p_2	... ...	p_n

Then

$$\sum_{i=1}^{n} p_i = p_1 + p_2 + \cdots + p_n = 1$$

The mean (μ) of the above probability distribution is defined as:

$$\mu = \frac{p_1x_1 + p_2x_2 + \cdots + p_nx_n}{p_1 + p_2 + \cdots + p_n} = \frac{\sum p_i x_i}{\sum p_i} = \sum p_i x_i$$

The variance (σ^2) is defined as:

$$\sigma^2 = \sum (x_i - \mu)^2 p_i = \sum (x_i{}^2 + \mu^2 - 2x_i\mu)p_i = \sum x_i{}^2 p_i + \mu^2 \sum p_i - 2\mu \sum x_i p_i$$

$$= \sum x_i{}^2 p_i + \mu^2(1) - 2\mu(\mu) = \sum x_i{}^2 p_i - \mu^2 = \sum x_i{}^2 p_i - \left(\sum p_i x_i\right)^2$$

Mean of a random variable X is also known as expected value and is denoted by E(X)

$$E(X) = \mu = p_1x_1 + p_2x_2 + \cdots + p_nx_n = \sum p_i x_i$$

$$\text{Variance } (\sigma^2) = E(X^2) - (E(X))^2$$

Example

A die is tossed twice. Getting a number greater than 4 is considered a success. Find the variance of the probability distribution of the number of success.

Solution:

Here p, probability of a number greater than 4=2/6=1/3 and q, probability of a number not greater than $4 = 1 - \frac{1}{3} = \frac{2}{3}$

$$P(X = 0) = q \times q = \frac{2}{3} \times \frac{2}{3} = \frac{4}{9}$$

$$P(X = 1) = p \times q + q \times p = \frac{1}{3} \times \frac{2}{3} + \frac{2}{3} \times \frac{1}{3} = \frac{4}{9}$$

$$P(X = 2) = p \times p = \frac{1}{3} \times \frac{1}{3} = \frac{1}{9}$$

Thus, we have:

x_i	p_i	$p_i x_i$	$x_i{}^2$	$p_i x_i{}^2$
0	4/9	0	0	0
1	4/9	4/9	1	4/9
2	1/9	2/9	4	4/9
Total		6/9		8/9

Hence, the mean $\mu = \sum p_i x_i = \frac{6}{9} = \frac{2}{3}$

The variance

$$\sigma^2 = \sum p_i x_i^2 - \mu^2 = \frac{8}{9} - \left(\frac{6}{9}\right)^2 = \frac{8}{9} - \frac{36}{81} = \frac{72-36}{81} = \frac{36}{81} = \frac{4}{9}$$

4.4 BINOMIAL DISTRIBUTION

Binomial distribution is a discrete probability distribution. This distribution was discovered by a Swiss Mathematician James Bernoulli (1654-1705). A Bernoullian trial is an experiment having only two possible outcomes i.e. success or failure. In other words the result of the trial are dichotomous e.g. in tossing of a coin either head or tail, the sex of a calf can be either male or female, a manufactured milk product or an engineering equipment or spare part will be either defective or non defective etc. This distribution can be used under the following conditions:

a) The random experiment is performed repeatedly a finite and fixed number of times i.e. n, the number of trials is finite and fixed.

b) The outcome of a trial results in the dichotomous classification of events i.e. each trial must result in two mutually exclusive outcomes, success or failure.

c) Probability of success (or failure) remains same in each trial i.e. in each trail the probability of success, denoted by p remains constant. q=1-p, is then termed as the probability of failure (non-occurrence).

d) Trials are independent i.e. the outcome of any trial does not affect the outcomes of the subsequent trials.

Theorem:

If **X** denotes the number of successes in n trials satisfying the above conditions, then **X** is a random variable which can take values 0,1,2,….,n i.e. no success, one success, two successes, ………, or all the **n** successes. The general expression for the probability of**r** successes is given by:$\mathbf{P(r) = P(X = r) = {}^nC_r p^r q^{n-r}}$ **for r=0,1,2,….,n**

Proof :

By the theorem of compound probability, the probability that r trials are success and the remaining (n-r) are failures in a sequence of n trials in a specified order say S,F,S,F,S,…..,S is given by

$$\begin{aligned} P(S \cap F \cap S \cap F \cap F \cap --- \cap S) &= P(S)P(F)P(S)P(F)P(F) --- P(S) \\ &= p.q.p.q.q --- p \\ &= (pxpxpx --- r \text{ times })x(qxqxq — (n-r)\text{times}) = p^r q^{(n-r)} \end{aligned}$$

But we are interested in any r trials being successes and since **r** trials can be chosen out of **n** trials in $\mathbf{{}^nC_r}$(mutually exclusive) ways. Therefore, by the theorem of total probability, the chance P (r) of **r** successes in a series of n independent trials is given by

$$\mathbf{P(r) = {}^nC_r p^r q^{n-r} \qquad 0 \le r \le n}$$

r can take only positive integer values.

Thus, the chance variate i.e. the number of successes, can take the values 0,1,2,..,r,..,n with corresponding probabilities
$q^n, {}^nC_1\, p\, q^{n-1}, .., {}^nC_r p^r q^{n-r}, .., p^n$

- The probability distribution of the number of successes so obtained is called the binomial probability distribution for the obvious reason that the probabilities are the various terms of the binomial expansion of **$(q+p)^n$.**

- The sum of probabilities

$$\sum_{r=0}^{n} p(r) = p(0) + p(1) + p(2) + --- p(r)$$
$$= q^n + {}^nC_1 pq^{n-1} + --- + {}^nC_r\, p^r q^{n-r} + --- + p^n = (q+p)^n$$
$$= 1$$

- The expression for P (X = r) is known as probability mass function of the Binomial distribution with parameter **n** and **p**. The random variable **X** following this probability law is called binomial variate with parmeter n and p denoted as **X~B(n,p).**Hence binomial distribution can be completely determined if n and p are known .

Example :

It is known that 40 % patients affected by tuberculosis die every year. 6 patients are admitted to a hospital suffering from tuberculosis. What is the probability that

(i) Three patients will die.
(ii) at least patients will die
(iii) all patients will be cured
(iv) no patients will be saved.

Solution

we have **p = 0.4 , q = 1- 0.40 = 0.6 and n=6**

In binomial distribution we have

$$\mathbf{P(r) = {}^nC_r .p^r .q^{n-r}}$$

(i) Prob. [Three patients will die]

$$P[r = 3] = P(3) = {}^6C_3 . (0.4)^3 (0.6)^3$$

$$P(3) = \frac{6!}{3!\,3!} (0.4)^3 (0.6)^3 = 20(0.4)^3(0.6)^3 = 0.2765$$

(ii) Prob. (at least five patients will die)

$$P(5) + P(6) = {}^6C_5 (0.4)^5 (0.6)^1 + {}^6C_6 (0.4)^6(0.6)^0$$
$$= 6 (0.4)^5 (0.6)^1 + (0.4)^6$$

= 0.0369 +0.0041= **0.0410**

(iii) Prob. (all patients will be cured) =1 - P (no patients will die)

$$1- P(0) =1 - {}^6C_0 (0.4)^0(0.6)^6$$
$$= 1 - (0.6)^6$$

= 1 - 0.0467 = *0.9533*

(iv) Prob. (no patients will be saved) = P (all patients will die)

= P(6)

= $^6C_6\,(0.4)^6\,(0.6)^0$

= $(0.4)^6$ =**0.0041**

Example of Binomial distribution

- The number of heads/tails in a sequence of coin flips
- Vote counts for two different candidates in an election
- The number of male/female employees in a company
- The number of accounts that are in compliance or not in compliance with an accounting procedure
- The number of successful sales calls
- The number of defective products in a production run

Properties of Binomial Distribution

i) **Mean of binomial distribution is np.**

Proof: First raw moment

$$\mu_1' = E(r) = \sum_{r=0}^{n} r\ n_{C_r} p^r q^{n-r} = \sum_{r=0}^{n} r \frac{n!}{r!(n-r)!} pp^{r-1}\, q^{n-r}$$

$$= \sum_{r=0}^{n} \frac{r.n.(n-1)!}{r.(r-1)![(r-1)-(r-1)]!} pp^{r-1} q^{(n-1)-(r-1)} = np \sum_{r=0}^{n} n-1_{C_{r-1}}\, p^{r-1} q^{(n-1)-(r-1)}$$

$$\sum_{r=0}^{n} np\, n-1_{C_{r-1}} p^{r-1} q^{(n-1)-(r-1)} =$$

$$np \sum_{r=0}^{n} n-1_{C_{r-1}} p^{r-1} q^{(n-1)-(r-1)} = np(q+p)^{n-1} = np$$

ii)**Variance of binomial distribution is npq**

Proof: Second raw moment

$$\mu_2' = E(r^2) = \sum_{r=0}^{n} r^2\, n_{C_r} p^r\, q^{(n-r)} = \sum_{r=0}^{n} \{r + r(r-1)\}\, n_{C_r} p^r q^{n-r}$$

$$= \sum_{r=0}^{n} r\, n_{C_r} p^r q^{n-r} + \sum_{r=0}^{n} r(r-1)\ n_{C_r} p^r q^{n-r} = np + n(n-1)p^2 = np + n^2p^2 - np^2$$

Variance

$$= \quad \mu_2 = \mu_2' - (\mu_1')^2 = np + n^2p^2 - np^2 - n^2p^2 = np(1-p) = npq$$

For the binomial distribution if mean and variance are known, we can arrive at the frequency distribution and variance is less than mean.

iii) The third and fourth central moment μ_3 and μ_4 can be obtained on the same lines.

$$\mu_3 = npq(q - p)$$
$$\mu_4 = npq[1 + 3(n - 2)pq]$$

iv) Pearson's constants β_1 & β_2 as well as γ_1 and γ_2 are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{[npq(1-2p)]^2}{(npq)^3} = \frac{[(1-2p)]^2}{npq} \qquad \gamma_1 = \sqrt{\beta_1} = \frac{(1-2p)}{\sqrt{npq}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{npq[1+3(n-2)pq]}{(npq)^2} = 3 + \frac{1-6pq}{npq}\ ,\ \gamma_2 = \beta_2 - 3 = \frac{1-6pq}{npq}$$

γ_1 shows that the binomial distribution is positively skewed if **$q > p$ or $p < 1/2$** and it is negatively skewed if **$q < p$ or $p > 1/2$** and it is symmetrical if **$p = q = 1/2$**.The binomial distribution is leptokurtic if **$pq < 1/6$** and latykurtic if **$pq > 1/6$.**

v) Mode of binomial distribution is determined by the value (n+1)p. If this value is an integer equal to k then the distribution is bi-modal, the two modal values being X=k and X=k-1.When this value is not an integer then the distribution has unique mode at $X=k_1$, the integral part of (n+1)p.

vi) Additive property: If X_1 is $B(n_1,p)$and X_2 is $B(n_2,p)$ and they are independent then their sum $X_1 + X_2$ is also a binomial variate $B(n_1 + n_2,p)$.

Example:

If the mean and variance of a Binomial Distribution are respectively 9 and 6, find the distribution.

Solution: Mean of Binomial Distribution is **np** and variance is **npq**

$$\therefore np = 9 \text{ and } npq = 6$$

$$Now\ \frac{npq}{np} = \frac{6}{9} \Rightarrow q = \frac{2}{3}$$

$$\therefore p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}$$

$$\therefore np = 9 \Rightarrow n.\frac{1}{3} = 9 \Rightarrow n = 3 \times 9 = 27$$

Hence, the Binomial Distribution is $\left(\frac{2}{3} + \frac{1}{3}\right)^{27}$

i.e. $^{27}C_r(1/3)^r(2/3)^{27-r}$

4.5 POISSON DISTRIBUTION

Poisson distribution is a limiting case of Binomial distribution under the following conditions:

- n, the no. of trials is indefinitely large i.e., $n \to \infty$
- p, the constant probability of success for each trial is indefinitely small i.e. $p \to 0$
- np= m (say) is finite. Thus, p = m/n, q = 1- m/n where m is a positive real number.

Under, the above three conditions the probability mass function of binomial distribution tends to the probability mass function of the Poisson distribution whose definition and derivation given below:

A random variable X is said to follow a Poisson distribution if it assumes only non-negative values and its probability mass function is given by

$$p(r, m) = P(x = m) = \frac{e^{-m} m^r}{r!}; \qquad r = 0,1,2,\dots\dots\dots,\infty$$

where m is known as the parameter of the distribution.

e = 2.7183(the base of the natural logarithm)

$$e^x = 1 + \frac{X}{1!} + \frac{X^2}{2!} + \frac{X^3}{3!} \dots\dots\dots\infty$$

$$e^{-x} = 1 - \frac{X}{1!} + \frac{X^2}{2!} - \frac{X^3}{3!} + \cdots (-1)^n \frac{X^r}{r!} + \cdots\infty$$

Proof: As $n \to \infty$ and np = m

p = m/n and q = 1-m/n

Probability function of binomial distribution is

$$P(r) = {}^nC_r\, p^r q^{n-r} = n!/r!(n-r)!\ p^r q^{n-r}$$

$$= \frac{n(n-1)(n-2)---[n-(r-1)]}{r!}\left(\frac{m}{n}\right)^r\left(1-\frac{m}{n}\right)^{n-r}$$

$$= \frac{m^r}{r!}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)--\left(1-\frac{r-1}{n}\right)x\left(1-\frac{m}{n}\right)^n\left(1-\frac{m}{n}\right)^{-r}$$

Taking limit as $n \to \infty$

$$= \frac{m^r}{r!}(1-0)(1-0)--(1-0)x\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n \lim_{n\to\infty}\left(1-\frac{m}{n}\right)^{-r}$$

We know that

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^n = e^{-m}$$

$$\lim_{n\to\infty}\left(1-\frac{m}{n}\right)^a = 1 \quad \text{a is not a function of n}$$

$$\text{Thus}, P(r) = \frac{m^r}{r!}\frac{e^{-m}.1}{.1} = \frac{e^{-m}m^r}{r!}; r = 0,1,2, ---\infty$$

Putting r = 0,1,2, ---- in above equation, we obtain the probabilities of r = 0,1,2, ---- successes respectively we get e^{-m}, $\frac{e^{-m}m^1}{1!}$, $\frac{e^{-m}m^2}{2!}$,---

Total probability is 1:

$$\sum_{r=0}^{\infty} p(r) = \sum_{r=0}^{\infty} P(x=r) = \sum_{r=0}^{\infty} \frac{e^{-m}m^r}{r!} = \sum_{r=0}^{\infty} P(x=r) = e^{-m} + me^{-m} + \frac{e^{-m}m^2}{2!} + ----$$

$$= e^{-m}\left(1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} \pm --\right) = e^{-m}\sum_{r=0}^{\infty}\frac{m^r}{r!} = e^{-m}.e^{m} = 1$$

If we know m, all the probabilities of the Poisson distribution can be obtained, therefore m is the only parameter of the Poisson distribution. The application of this distribution in solving problems is illustrated through following examples.

Example

A manufacturer of screws knows that 5% of his product is defective. If he sells his product in a carton of 100 items and guarantees that not more than 10 items will be defective. What is the probability that the carton will fail to meet the guaranteed quality?

Solution:

In this example p = 0.05, n = 100. Therefore, m = n. p = 100 (0.05) = 5

Prob. [That the carton will fail to meet the guaranteed quality] = 1- Prob. [The carton will meet the guaranteed quality] = Prob. [Not more than 10 items will be defective] = 1 - P [r ≤ 10]
= 1- [P(0) + P(1) + P(2) + P(3)+…………+ P(10)]

In case of Poisson distribution $P(r) = \frac{e^{-m}m^r}{r!}$

Therefore, we have

P(r > 10) = 1- P(r ≤ 10) = 1 - $\left(\frac{e^{-5}5^0}{0!} + \frac{e^{-5}5^1}{1!} + \frac{e^{-5}5^2}{2!} + \cdots + \frac{e^{-10}5^{10}}{10!}\right)$

$= 1 - e^{-5}\left[1 + 5 + \frac{5^2}{2!} + \frac{5^3}{3!} + \cdots + \frac{5^{10}}{10!}\right] = 1 - 0.9865 = 0.0135$

Examples of Poisson Distribution

- The hourly number of customers arriving at a bank

- The daily number of accidents on a particular stretch of highway
- The hourly number of accesses to a particular web server
- The daily number of emergency calls in Dallas
- The number of typos in a book
- The monthly number of employees who had an absence in a large company
- Monthly demands for a particular product

Properties of Poisson Distribution

i) Mean of the Poisson distribution is m

$$\mu_1' = \text{Mean} = \sum_{r=0}^{\infty} r\frac{e^{-m}m^r}{r!} = \sum_{r=0}^{\infty} r\frac{e^{-m}m^r}{(r-1)!} = m\sum_{r=0}^{\infty}\frac{e^{-m}m^{r-1}}{(r-1)!}$$

$$= me^{-m}\left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + ---\right] = m\,e^{-m}e^{m} = m$$

ii) Variance of the Poisson distribution is

$$\text{Variance} = \sum_{r=0}^{\infty} r^2p(r) - \left(\sum_{r=0}^{\infty} r\,p(r)\right)^2 = \sum_{r=0}^{\infty} r^2p(r) - (m)^2$$

where,

$$\mu_2' = \sum_{r=0}^{\infty} r^2\frac{e^{-m}m^r}{r!} = \sum_{r=0}^{\infty}[r + r(r-1)]\frac{e^{-m}m^r}{r!} = \sum_{r=0}^{\infty} r\frac{e^{-m}m^r}{r!} + \sum_{r=0}^{\infty} r(r-1)\frac{e^{-m}m^r}{r!}$$

$$= m + e^{-m}m^2\sum_{r=0}^{\infty}\frac{m^{r-2}}{(r-2)!} = m + e^{-m}m^2\left[1 + m + \frac{m^2}{2!} + \frac{m^3}{3!} + ---\right]$$

$$= m + e^{-m}m^2e^{m} = m+m^2$$

$$\text{Variance} = \mu_2' - (\mu_1')^2 = m+m^2 - (m)^2 = m$$

Hence, for Poisson distribution with parameter m mean is equal to variance.

iii) Third and fourth central moments μ_3 and μ_4

$$\mu_3 = m, \quad \mu_4 = 3m^2 + m$$

iv) Pearson's constants β_1 & β_2 as well as γ_1 and γ_2 are given by

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(m)^2}{(m)^3} = \frac{1}{m},\ \gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3m^2+m}{(m)^2} = 3 + \frac{1}{m},\quad \gamma_2 = \beta_1 - 3 = \frac{1}{m}$$

It may be noted that the first three central moments of the Poisson distribution are identical and are equal to the value of parameter itself

namely 'm'. Hence Poisson distribution is always a positively skewed distribution as m>0 as well as leptokurtic. As the value of m increases γ_1 decreases and the thus Skewness is reduced for increasing values of m. As $m\to\infty$, γ_1 and γ_2 tend to zero. So we conclude that as $m\to\infty$, the curve of the Poisson distribution tends to be symmetrical curve for large values of m.

v) Mode of Poisson distribution is determined by the value m. If m is an integer then the distribution is bi-modal, the two modal values being X=m and X=m-1.When m is not an integer then the distribution has unique modal value being integral part of m.

vi) Additive property: If X_1 and X_2 are two independent Poisson variate with parameters m_1 and m_2 then their sum $X_1 + X_2$ is also a Poisson variate with parameter $m_1 + m_2$.

Example

The mean of the Poisson distribution is 2.25. Find the other constants of the distribution.

Solution:

We have $m = 2.25$

$$\sigma = \sqrt{m} = \sqrt{2.25} = 1.5$$

$$\mu_1 = 0$$

$$\mu_2 = m = 2.25$$

$$\mu_3 = m = 2.25$$

$$\mu_4 = m + 3m^2 = 2.25 + 3(2.25)^2 = 2.25 + 15.1875 = 17.4375$$

$$\beta_1 = \frac{1}{m} = \frac{1}{2.25} = 0.444$$

$$\beta_2 = 3 + \frac{1}{m} = 3 + 0.444 = 3.444$$

$$\gamma_1 = \sqrt{\beta_1} = \frac{1}{\sqrt{m}} = \frac{1}{1.5} = 0.67$$

$$\gamma_2 = \beta_2 - 3 = 3 + \frac{1}{m} - 3 = \frac{1}{2.25} = 0.444$$

This curve is positively Skewed and Leptokurtic.

CHECK YOUR PROGRESS - 1

1. List the types of random variable
2. What are the properties of Binominal distribution?
3. List out few example of Poisson distribution

4.6 NORMAL DISTRIBUTION

Normal distribution is one of the important distribution in

continuous probability distribution. Normal distribution is probably the most important and widely used theoretical distribution. Normal distribution unlike the Binomial and Poisson is a continuous probability distribution. It has been observed that a vast number of variables arising in studies of agricultural and dairying, social, psychological and economic phenomena tend to follow normal distribution. The normal distribution was first discovered by French Mathematician Abraham De-Moivre in 1733, who obtained this continuous distribution as a limiting case of the Binomial distribution. But it was later rediscovered and applied by Laplace and Karl Gauss. It is also known as Gaussian distribution after the name of Karl Friedrich Gauss.

A continuous random variable X is said to have a normal distribution with parameters μ (mean) and σ (standard deviation), if its density function is given by the probability law

$$f(x) = \frac{1}{\sqrt{2\pi}.\sigma}.e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

where π and e are given by π = 22/7 and e=2.7183 (base of natural logarithms).

Properties of normal distribution:

1) A random variable X with mean μ and variance σ^2 following the normal law given above is represented as $X \sim N(\mu, \sigma^2)$.
2) If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma}$, Z is defined as a standard normal variate with E(Z)=0 and Var (Z)=1and we write Z~N(0, 1)
3) The p.d.f. of a standard normal variate Z is given by

$$\emptyset(z) = \frac{1}{\sqrt{2\pi}}.e^{-\frac{z^2}{2}}, -\infty < z < \infty, where\ Z = \frac{X-\mu}{\sigma}$$

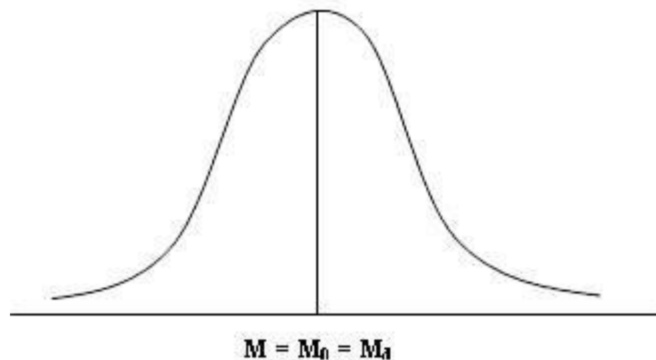
4) Normal distribution is a limiting form of the binomial distribution when
   a) n, the number of trials is indefinite large, i.e. n→∞ and
   b) neither p nor q is very small.
5) Normal distribution is a limiting form of Poisson distribution when its mean m is large and n is also large.

Characteristics of Normal Distribution

It has the following properties:

1. The graph of f(x) is bell shaped unimodal and symmetric curve as

shown in the Fig. 12.1. The top of the bell is directly above the mean (μ).



Normal probability curve

2. The curve is symmetrical about the line X = μ, (Z = 0) i.e., it has the same shape on either side of the line X = μ, (or Z = 0). This is because the equation of the curve Ø(z) remains unchanged if we change z to -z.
3. Since the distribution is symmetrical, mean, median and mode coincide. Thus, Mean = Median = Mode = μ
4. Since Mean = Median = Mode = μ, the ordinate at X = μ, (Z = 0) divides the whole area into two equal parts. Further, since total area under normal probability curve is 1, the area to the right of the ordinate as well as to the left of the ordinate at X = μ (or Z = 0) is 0.5
5. Also, by virtue of symmetry the quartiles are equidistant from median (μ), i.e,

$$Q_3 - M_d = M_d - Q_1$$

6. Since the distribution is symmetrical, all moments of odd order about the mean are zero. Thus $\mu_{2n+1} = 0$; (n = 0,1,2,………) i.e. $\mu_1 = \mu_3 = \mu_5 = --- = 0$.
7. The moments (about mean) of even order are given by

$$\mu_{2n} = 1.3.5 \ldots (2n-1)\sigma^{2n}, (n = 1,2,3 \ldots)$$

Putting n=1 and 2 we get

$\mu_2 = \sigma^2$ and $\mu_4 = 3\sigma^4$

$$\therefore \beta_1 = \frac{\mu_3^2}{\mu_2^3} = 0$$

and

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3\sigma^4}{\sigma^4} = 3$$

8. Since the distribution is symmetrical, the moment coefficient of skewness based on moments is given by $\beta_1 = 0 \Rightarrow \gamma_1 = 0$
9. The coefficient of kurtosis is given by $\beta_2 = 3 \Rightarrow \gamma_2 = 0$
10. No portion of the curve lies below the x-axis, since f(x) being the probability can neverbe negative.

11. Theoretically, the range of the distribution is from $-\infty <$ to $< \infty$. But practically, range = 6σ
12. As x increases numerically [i.e. on either side of $X = \mu$], the value of f(x) decreases rapidly, the maximum probability occurring at $X = \mu$ and is given by

$$[f(x)]_{max} = \frac{1}{\sqrt{2\pi}\sigma}$$

Thus maximum value of f(x) is inversely proportional to the standard deviation. For large values of σ, f(x) increases, i.e., the curve has a normal peak.

13. Distribution is unimodal with the only mode occurring at $X = \mu$.
14. X-axis is an asymptote to the curve i.e., for numerically large value of X (on either side of the line ($X = \mu$)), the curve becomes parallel to the X-axis and is supposed to meet it at infinity.
15. A linear combination of independent normal variates is also a normal variate. If $X_1, X_2, \ldots, X_n$ are independent normal variates with mean $\mu_1, \mu_2, \ldots, \mu_n$ and standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_n$ respectively then their linear combination

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

Where $a_1, a_2, \ldots, a_n$ are constants, is also a normal variate with Mean = $a_1\mu_1 + a_2\mu_2 + \ldots + a_n\mu_n$ and Variance = $a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ldots + a_n^2\sigma_n^2$. In particular, if we take $a_1 = a_2 = \ldots = a_n = 1$ then we get $X_1 + X_2 + \ldots + X_n$ is a normal variate with mean $\mu_1 + \mu_2 + \ldots + \mu_n$ and variance $\sigma_1^2 + \sigma_2^2 + \ldots + \sigma_n^2$. Thus, the sum of independent normal variates is also a normal variate with mean equal to sum of their means and standard deviation equal to square root of sum of the squares of their standard deviations. This is known as the Re-productive or Additive Property of the Normal distribution.

16. Mean Deviation (M.D.) about mean or median or mode is given by

$$\text{M.D.} = \sqrt{\frac{2}{\pi}}.\sigma \cong \frac{4}{5}\sigma$$

17. Quartiles are given (in terms of μ and σ) by

$$Q_1 = \mu - 0.6745\sigma \text{ and } Q_3 = \mu + 0.6745\sigma$$

18. Quartile deviation (Q.D.) is given by

$$\text{Q.D.} = \frac{Q_3 - Q_1}{2} = 0.6745\sigma \cong \frac{2}{3}\sigma$$ Also

$$\text{Q.D.} = \frac{2}{3}\sigma = \frac{4}{6}\sigma = \frac{5}{6} \times \frac{4}{5}\sigma = \frac{5}{6}\text{ M.D.}$$

$$\therefore \text{Q.D.} = \frac{5}{6}\text{ M.D.}$$

19. We have (approximately):

$$Q.D.:M.D:S.D.::\frac{2}{3}\sigma:\frac{4}{5}\sigma:\sigma::\frac{2}{3}:\frac{4}{5}:1 \Rightarrow Q.D.:M.D.:S.D.::10:12:15$$

From property 18 we also have $4S.D. = 5M.D. = 6Q.D.$

20. Points of inflexion of the normal curve are at X = μ ± σ i.e. they are equidistant from mean at a distance of σ and are given by :

$$X = \mu \pm \sigma, \quad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-1/2}$$

21. Area property: One of the most fundamental property of the normal probability curve is the area property. If $X \backsim N(\mu, \sigma^2)$, then the probability that random value of X will lie between X= μ and X= x_1 is given

$$P(\mu < X < x_1) = \int_{\mu}^{x_1} f(x).dx = \frac{1}{\sqrt{2\pi}\,\sigma}\int_{\mu}^{x_1} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

put $z = \frac{x-\mu}{\sigma} \Rightarrow x = \mu + \sigma z$; ∴ at $x = \mu, z = 0$; and at $x = x_1, z = \frac{x_1 - \mu}{\sigma}$

$= z_1$

∴ $P(0 < x < x_1) = P(0 < z < z_1)$

$$= \frac{1}{\sqrt{2\pi}}\int_0^{z_1} e^{-\frac{1}{2}z^2} dz = \int_0^{z_1} \emptyset(z)dz$$

Where $\emptyset(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$, is the probability function of standard normal variate. The definite integral $\int_0^{z_1} \emptyset(z)dz$, is known as **Normal Probability integral** and gives the area under standard normal curve between the ordinate z=0 and z = z_1. These areas have been provided in the form of table for different values of z_1 at the intervals of 0.01 which are available in any standard text books of statistics.

Particular Cases:

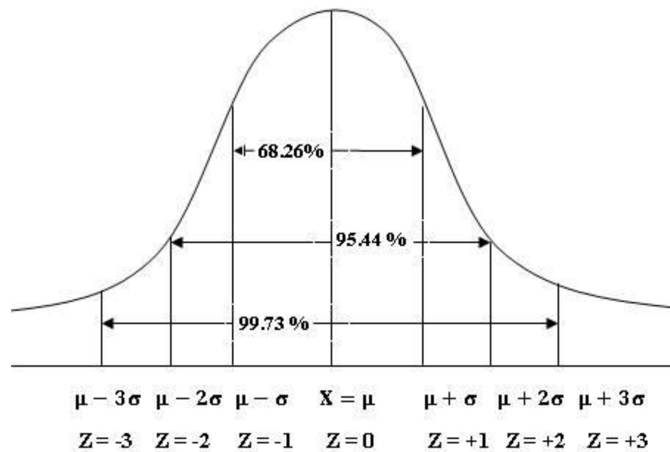
1. In particular, the probability that a random variable X lies in the interval (μ-σ, μ+σ) is given by

$$P(\mu - \sigma < X < \mu + \sigma) = \int_{\mu-\sigma}^{\mu+\sigma} f(x)dx.$$

$$P(-1 < Z < 1) = \int_{-1}^{1} \emptyset(z)dz = 2\int_0^1 \emptyset(z)dz = 2(0.3413) = 0.6826$$

The area under the normal probability curve between the ordinates at X= μ-σ and X= μ+σ is 0.6826.In other words, the range X = μ-σ covers 68.26% of the observations (as shown in Fig). This is known as 1σ limit of normal distribution

This is known as 1σ limit of normal distribution



1σ, 2σ and 3σ under Normal Probability Curve

2. The probability that random variable X lies in the interval (μ-2σ, μ+2σ) is given by

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = \int_{\mu-2\sigma}^{\mu+2\sigma} f(x)dz. \Rightarrow P(-2 < Z < 2) = \int_{-2}^{2} \emptyset\,(z)dz$$

$$= 2\int_{0}^{2} \emptyset\,(z)dz = 2(0.47725) = 0.95445$$

The area under the normal probability curve between the ordinates at X= μ-2σ and X= μ+2σ is 0.95445. In other words, the range X = μ+2σ covers 95.445% of the observations (as shown in Fig.). This is known as 2σ limits of normal distribution and is considered as warning limit in case of statistical quality control which implies that it is a warning to the manufacturer that the manufacturing process is going out of control.

3. The probability that random variable X lies in the interval (μ−3σ, μ+3σ) is given by

$$P(\mu - 3\sigma < X < \mu + 3\sigma = P(-3 < Z < 3)$$

$$\int_{-3}^{3} \emptyset\,(z)dz = 2\int_{0}^{3} \emptyset\,(z)dz = 2(0.49865) = 0.9973$$

The area under the normal probability curve between the ordinates at X=μ−3σ and X= μ+3σ is 0.9973.In other words, the range X = μ+2σ covers 99.73% of the observations (as shown in Fig.). This is known as 3σ limits of normal distribution and it implies the manufacturing process is out of control in case of statistical quality control.

Thus, the probability that a normal variate X lies outside the range μ- 3σ is given as

$$P(|X - \mu| > 3\sigma) = P(|z| > 3) = 1 - P(-3 < z < 3) = 1 - 0.9973 = 0.0027$$

Thus, in all probability, we should expect a normal variate to lie within the range μ- 3σ though theoretically may range from −∞ to ∞.

Example

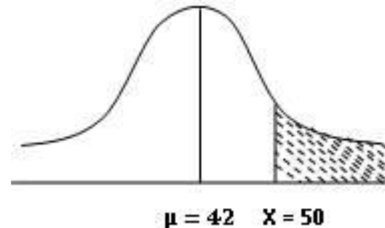
An Intelligence test was administrated to 1000 students. The average score of students was 42 with standard deviation of 24. Find

(a) Number of students exceeding a score of 50
(b) Number of students scoring between 30 & 58
(c) Value of score exceeded by top 100 students.

Solution:

In this problem μ =42 and σ = 24 and let X denote the score obtained

(a) Number of students exceeding score 50



As shown in figure we want to find

P(X>50) i.e. probability of shaded portion

At X=50, $Z = \frac{50-42}{24} = \frac{8}{24} = 0.334$

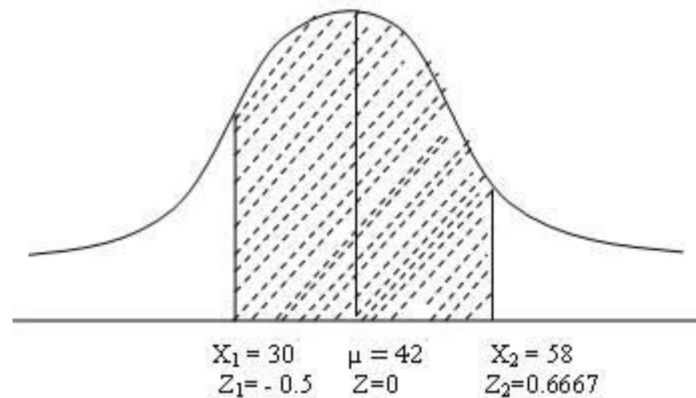
$P(X>50) = P(Z > 0.334) = 0.5 - P(0 \leq Z \leq 0.334) = 0.5 - 0.1308 = \mathbf{0.3692}$

No of students = 1000 * 0.3692= 369.2 ~ 369 students

(b) Number of students scoring between 30 and 58

As shown in figure we want to find

P(30<X<58) i.e. probability of shaded portion



At $X_1 = 30 \quad Z_1 = \frac{30-42}{24} = -0.5$

$P(Z_1 > -0.5) = P(0 \leq Z_1 \leq 0.5) = \mathbf{0.1915}$

At $X_2 = 58 \quad Z_2 = \frac{58-42}{24} = 0.6667$

$P(Z_2 < 0.6667) = P(0 \leq Z_2 \leq 0.6667) = \mathbf{0.2476}$

$P(30<X<58)=P(-0.5\leq Z\leq 0.6667) = 0.1915+0.2476 =$ **0.4391**

No of students = 1000 * .4391 = 439.1 ~ 439 students

(c) Value of score exceeded by top 100 students.

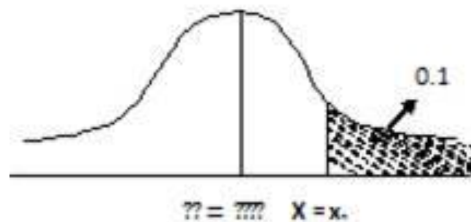
Let x_1 be the value of score exceeded by top 100 students, the probability of top 100 students = 100/N = 100/1000 = 0.1 such that $P(X>x_1) = 0.1$

At $X = x_1$, $Z = \frac{x_1 - 42}{24} = Z_1$.

From Fig the $P(X>x_1)$ shown as shaded region

$P(X>x_1) = P(Z>Z_1) = 0.1 \Rightarrow P(0 \leq Z \leq Z_1) = 0.4 \Longrightarrow \frac{x_1-42}{24}$ **= 1.286**

x_1= 72.86 ~73



Conclusion

(a) **369 students scored more than 50.**

(b) 439 students scored between 30 & 58.

(c) Minimum score of top 100 students is 73.

4.7 SUMMARY

- Conditions for the binomial probability distribution are
  - i. The trials are independent
  - ii. The number of trials is finite
  - iii. Each trial has only two possible outcomes called success and failure.
  - iv. The probability of success in each trial is a constant.
- The parameters of the binomial distributions are *n* and *p*
- The mean of the binomial distribution is *np* and variance are *npq*
- Poisson distribution as limiting form of binomial distribution when n is large, *p* is small and *np*is finite.
- The Poisson probability distribution is $x = 0,1,2,3\ldots$ Where $\lambda =$

np

- The mean and variance of the Poisson distribution is λ.
- The λ is the only parameter of Poisson distribution.
- Poisson distribution can never be symmetrical.
- It is a distribution for rare events.
- Normal distribution is the limiting form of binomial distribution when *n* is large and neither *p n*or *q* is small

4.8 KEY WORDS

Random Variables, Binomial Distibution, Poisson Distirbution, Normal Distribution

4.9 ANSWER TO CHECK YOUR PROGRESS

1. **Discrete random variable, continuous random variable,**
2. **Mean = np, SD = $\sqrt{npq}$, variance = npq**
3. **Examples**
    - The hourly number of customers arriving at a bank
    - The daily number of accidents on a particular stretch of highway
    - The hourly number of accesses to a particular web server
    - The daily number of emergency calls in 108
    - The number of types in a book
    - The monthly number of employees who had an absence in a large company

4.10 QUESTIONS AND EXERCIS

SHORT ANSWER QUESTIONS:

1. Define Binomial distribution
2. Mention the properties of Normal distribution
3. What are the main characteristics of Poisson distribution
4. Determine the binomial distribution for which the mean is 4 and variance 3. Also find P(X = 15)

LONG ANSWER QUESTIONS

1. What is meant by probability distribution of a discrete random

variable?

2. Define Binomial distribution? what are the main characteristics of binomial distribution

3. Write the main characteristics of normal distribution

4. Fit the Poisson distribution to the following

X	0	1	2	3	4	5
f	120	82	52	22	4	0

4.11 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers and Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing Company Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw Hill Publishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., NewDelhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., New Delhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. Prentice Hall of India Pvt. Ltd., New Delhi.

UNIT – V ESTIMATION

Structure

5.1. INTRODUCTION

The sampling process is used to draw statistical inference about the characteristics of a population or process of interest. On many occasions we do not have enough information to calculate an exact value of population parameters (such as μ, σ and p) and therefore make the best estimate of this value from the corresponding sample statistics (such as x , s, and P). The need to use the sample statistic to draw conclusions about the population characteristic is one of the fundamental applications of statistical inference in business and economics.

5.2 Reasons for Making Estimates

A few applications of statistical estimation are given below :

A production manager needs to determine the proportion of items being manufactured that do not match with quality standards.

A mobile phone service company may be interested to know the average length of a long distance telephone call and its standard deviation

A bank needs to understand consumer awareness of its services and credit schemes.

Any service centre needs to determine the average amount of time a customer spends in queue.

In all such cases, a decision-maker needs to examine the following two concepts that are useful for drawing statistical inference about an unknown population or process parameters based upon random samples:

Estimation– a sample statistic to estimate an unknown parameter value

Hypothesis testing– a claim or belief about an unknown parameter value.

In this lesson we shall discuss methods to estimate unknown population parameter and then to determine the range of values (confidence interval) likely to contain the parameter value.

5.3 TYPES OF ESTIMATES

Let us first know the concept of 'estimate' as used in Statistics. According to some dictionaries, an estimate is a valuation based on opinion or roughly made from imperfect or incomplete data. This definition may apply, for example, when an individual who has an opinion about the competence of one of his colleagues. But, in Statistics the term estimate is not used in this sense. In Statistics too the estimates are made when the information available is incomplete or imperfect. However, such estimates are made only when they are based on sound judgement or experience and when the samples are scientifically selected.

There are two types of estimates that we can make about a population : *a point estimate* and an *interval estimate.* A point estimate is a single number, which is used to estimate an unknown population parameter. Although a point estimate may be the most common way of expressing an estimate, it suffers from a major limitation since it fails to indicate how close it is to the quantity it is supposed to estimate. In other words, a point estimate does not give any idea about the reliability of precision of the method of estimation used. For instance, if someone claims that 40 percent of all children in a certain town do not go to the school and are devoid of education, it would not be very helpful if this claim is based on a small number of households, say, 20. However, as the number of households interviewed for this purpose increases from 20 to 100, 500 or even 5,000, the claim that 40 percent of children have no school education would become more and more meaningful and reliable. This makes it clear that a point estimate should always be accompanied by some relevant information so that it is possible to judge how far it is reliable.

The second type of estimate is known as the interval estimate. It is a range of values used to estimate an unknown population parameter. In case of an interval estimate, the error is indicated in two ways: first by the extent of its range; and second, by the probability of the true population parameter lying within that range. Taking our previous example of 40 percent children not having a school education, the statistician may say that actual percentage of such children in that town may lie between 35 percent and 45 percent. Thus, he will have a better idea of the reliability of such an estimate as compared to the point estimate of 40 percent.

Estimator and Estimate

When we make an estimate of a population parameter, we use a sample statistic. This sample statistic is an estimator.

For example, the samples mean $x = \frac{\sum_{i=1} xi}{n}$

is a point estimator of the population mean μ. The value obtained by the estimator is known as an estimate. Many different Statistics can be used to estimate the same parameter. For example, we may use the sample mean or the sample median or even the range to estimate the population mean. The

question here is: how can we evaluate the properties of these estimates, compare then with one another, and finally, decide which the 'best' is? The answer to this question is possible only when we have certain criteria that a good estimator must satisfy. These criteria are briefly discussed below.

5.4 POINT ESTIMATION

In point estimation, a single sample statistic (such as $\bar{x}$, *s*, and $\bar{p}$) is calculated from the sample to provide a best estimate of the true value of the corresponding population parameter (such as μ, σ and $\bar{p}$). Such a single relevant statistic is termed as *point estimator,* and the

value of the statistic is termed as *point estimate.* For example, we may calculate that 10 per cent of the items in a random sample taken from a day's production are defective. The result '10 per cent' is a point estimate of the percentage of items in the whole lot that are defective. Thus, until the next sample of items is not drawn and examined, we may proceed on manufacturing on the assumption that any day's production contains 10 per cent defective items.

5.5 INTERVAL ESTIMATION

Generally, a point estimate does not provide information about 'how close is the estimate' to the population parameter unless accompanied by a statement of possible sampling errors involved based on the sampling distribution of the statistic. It is therefore important to know the precision of an estimate before relying on it to make a decision. Thus, decision - makers prefer to use an *interval estimate* that is likely to contain the population parameter value. However, it is also important to state 'how confident' he is that the interval estimate actually contains the parameter value. Hence an interval estimate of a population parameter is therefore a *confidence interval* with a statement of confidence that the interval contains the parameter value.

5.6 CRITERIA OF A GOOD ESTIMATOR

There are four criteria by which we can evaluate the quality of a statistic as an estimator.

These are: unbiasedness, efficiency, consistency and sufficiency.

5.7 Unbiasedness

This is a very important property that an estimator should possess. If we take all possible samples of the same size from a population and calculate their means, the mean μ x of all these means will be equal to the mean μ of the population. This means that the sample mean is an unbiased estimator of the population mean μ. When the expected value (or mean) of a sample statistic is equal to the value of the corresponding population parameter, the sample statistic is said to be an unbiased estimator.

Suppose we take the smallest sample observation as an estimator of the population mean μ, it can be easily shown that this estimator is biased. Since the smallest observation must be less than the mean, its expected value must be less than μ. Symbolically, $E(Xs) < \mu$, where Xs stands for

the smallest item and E stands for the expected value. Thus, this estimator is biased downwards. The extent of bias is the difference between the expected value of the estimator and the value of the parameter. In this case, bias is equal to E(Xs)- μ. In contrast, the biases for the sample mean x is zero.

5.8 Consistency

Another important characteristic that an estimator should possess is consistency. Let us take the case of the standard deviation of the sampling distribution of $\bar{x}$. The standard deviation of the sampling distribution of sample mean is computed by following formula :

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

The formula states that the standard deviation of the sampling distribution of $\bar{x}$ decreases as the sample size increases and *vice versa*. When the sample size *n* increases, the population standard deviation σ is to be divided by a higher denominator. This results in the reduced value of sample standard deviation σ_{ξ}. Let us take an example.

Illustration 13.1: ***A company has 4,000 employees whose average monthly wage comes to Rs.4,800 with a standard deviation of Rs.1,200. Let $\bar{x}$ be the mean monthly wage for a random sample of certain employees selected from this company. Find the mean and standard deviation of ξ for a sample size of (a) 81, (b) 100 and (c) 180.***

Solution

From the given information, for the population of all employees, N = 4,000 μ = Rs.4,800 σ = Rs.1,200.

The mean μ_{ξ} of the sampling distribution of the ξ is $\mu_{\xi} = \mu$ = Rs.4,800.

As n = 81 and N = 4,000, which gives n/N = 0.01. At this value is less than 0.05, the standard deviation of ξ is obtained by using the formula. Substituting the values.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or, } \sigma_{\bar{x}} = \frac{1,200}{\sqrt{81}} = \frac{1,200}{9} = \text{Rs.}133.33$$

In this case, n = 100 and n/N = 100/4,000 = 0.025, which is also less than 0.05. The mean and the standard deviation ξ are

$\mu_{\xi} = \mu$ = Rs.4,800

$$\sigma_{x} = \frac{\sigma}{n} \text{ or, } \sigma_{x} = \frac{1,200}{100} = \frac{1,200}{10} = \text{Rs.}120$$

In this case, n = 180 and n/N = 180/4,000 = 0.045, which again is less than 0.05. The mean and the standard deviation ξ are

$\mu_{\bar{x}} = \mu$ = Rs.4,800

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \text{ or, } \sigma_{\bar{x}} = \frac{1,200}{\sqrt{180}} = \frac{1,200}{13.42} = \text{Rs.}89.42$$

From the above three sets of calculation, it becomes clear that the mean of the sampling distribution of $\bar{x}$ is always equal to the mean of the population regardless of the sample size. But, in case of the standard deviation, we find the change. In the given example, we find that standard deviation of $\bar{x}$ decreased from Rs.189.87 to Rs.120 and then to Rs.133.33 as the sample size increased from 40 to 100 and then to 180.

5.9 Efficiency

Another desirable property of a good estimator is that it should be efficient. Efficiency is measured in terms of size of the standard error of the statistic. Since an estimator is a random variable, it is necessarily characterised by a certain amount of variability. This means that some estimates may be more variable than others. Just as bias is related to the expected value of the estimator, so efficiency can be defined in terms of the variance. In large samples, for example, the variance of the sample mean is $V(\bar{x})=\sigma^2/n$. As the sample size n increases, the variance of the sample mean ($V\bar{x}$) becomes smaller, so the estimator becomes more efficient. This criterion, when applied to large samples, gives better estimates as compared to the small ones.

The efficiency of one estimator in relation to another estimator can be judged by comparing their sampling variances. Thus, efficiency relates to the size of the standard error. Given the same sample size, the statistic that has a smaller standard error is preferable as it is efficient in relation to another statistic that has a larger standard error. The sampling distribution of the mean and the median have the same mean, that is, the population mean. However, the variance of the sampling distribution of the means is smaller than the variance of the sampling distribution of the medians. As such, the sample mean is an efficient estimator of the population mean, while the sample median is an inefficient estimator.

5.10 Sufficiency

The fourth property of a good estimator is that it should be sufficient. A sufficient statistic utilises all the information a sample contains about the parameter to be estimated. ξ, for example, is a sufficient estimator of the population mean μ. It implies that no other estimator of μ, such as the sample median, can provide any additional information about the parameter μ. Likewise, we can say that the sample proportion π.

Having looked into properties of a good estimator briefly, a pertinent question arises: how can we find estimators with these desirable properties? This brings us to the method of maximum likelihood

5.6.2 METHOD OF MAXIMUM LIKELIHOOD (ML)

The maximum likelihood method provides estimators with the desirable properties such as efficiency, consistency and sufficiency, which we have just discussed. It usually does not give an unbiased estimate. Let us take an example to explain this method.

Example: ***Suppose we want to estimate the average grade*** **μ** ***of a large number of students. A random sample of size*** **n** ***= 64 is taken and the sample mean*** $\bar{x}$ ***is found to be 90 marks. Now, the assumption on which we have to base our reasoning is that the random sample of*** **n** ***= 64 is representative of the population. We saw how samples that were similar to the population had greater probability of being selected.***

Let us now reverse this reasoning as follows: we have before us a random sample size n = 64 and $\bar{x}$ = 90 marks. From which population did it most probably come-a population with μ = 85, 90 or 95? According to our earlier approach, we would think that it most probably came from a population with μ =90 marks. Thus, it can be concluded that the population mean μ, based on our sample, is most likely to be μ=90 marks.

A point worth noting is that the population mean μ is either 90 or not; it has only one value.

Hence, we have used the term *likely* instead of probably.

This technique to find the estimators was first used and developed by Sir R.A. Fisher in 1922, who called it the maximum likelihood method

5.11 Confidence Intervals

There are two types of estimates for each population parameter: the point estimate and confidence interval (CI) estimate. For both continuous variables (e.g., population mean) and dichotomous variables (e.g., population proportion) one first computes the point estimate from a sample. Recall that sample means and sample proportions are unbiased estimates of the corresponding population parameters.

For both continuous and dichotomous variables, the confidence interval estimate (CI) is a range of likely values for the population parameter based on:

the point estimate, e.g., the sample mean
the investigator's desired level of confidence (most commonly 95%, but any level between 0-100% can be selected)
and the sampling variability or the standard error of the point estimate.

Strictly speaking a 95% confidence interval means that if we were to take 100 different samples and compute a 95% confidence interval for each sample, then approximately 95 of the 100 confidence intervals will contain the true mean value (μ). In practice, however, we select one random sample and generate one confidence interval, which may or may not contain the true mean. The observed interval may over- or underestimate μ. Consequently, the 95% CI is the likely range of the true, unknown parameter. The confidence interval does not reflect the variability in the unknown parameter. Rather, it reflects the amount of random error in the sample and provides a range of values that are likely to include the unknown parameter. Another way of thinking about a confidence interval is that it is the range of likely values of the parameter (defined as the point estimate + margin of error) with a specified level of confidence (which is similar to a probability).

Suppose we want to generate a 95% confidence interval estimate for an unknown population mean. This means that there is a 95% probability that the confidence interval will contain the true population mean. Thus, P([sample mean] - margin of error < μ < [sample mean] + margin of error) = 0.95.

The Central Limit Theorem introduced in the module on Probability stated that, for large samples, the distribution of the sample means is approximately normally distributed with a mean:and a standard deviation (also called the standard error):

For the standard normal distribution, $P(-1.96 < Z < 1.96) = 0.95$, i.e., there is a 95% probability that a standard normal variable, Z, will fall between -1.96 and 1.96. The Central Limit Theorem states that for large samples:

By substituting the expression on the right side of the equation:

Using algebra, we can rework this inequality such that the mean (μ) is the middle term, as shown below .then and finally

This last expression, then, provides the 95% confidence interval for the population mean, and this can also be expressed as:

Thus, the margin of error is 1.96 times the standard error (the standard deviation of the point estimate from the sample), and 1.96 reflects the fact that a 95% confidence level was selected. So, the general form of a confidence interval is:

point estimate + Z SE (point estimate)

where Z is the value from the standard normal distribution for the selected confidence level (e.g., for a 95% confidence level, Z=1.96).

In practice, we often do not know the value of the population standard deviation (σ). However, if the sample size is large ($n > 30$), then the sample standard deviations can be used to estimate the population standard deviation.

512 Determining the Sample Size in Estimation

Sample size determination is the act of choosing the number of observations or replicates to include in a statistical sample. The sample size is an important feature of any empirical study in which the goal is to make inferences about a population from a sample. In practice, the sample size used in a study is usually determined based on the cost, time, or convenience of collecting the data, and the need for it to offer sufficient statistical power. In complicated studies there may be several different sample sizes: for example, in a stratified survey there would be different sizes for each stratum. In a census, data is sought for an entire population, hence the intended sample size is equal to the population. In experimental design, where a study may be divided into different treatment groups, there may be different sample sizes for each group.

Sample sizes may be chosen in several ways:

using experience – small samples, though sometimes unavoidable, can result in wide confidence intervals and risk of errors in statistical hypothesis testing.
using a target variance for an estimate to be derived from the sample eventually obtained, i.e. if a high precision is required (narrow confidence interval) this translates to a low target variance of the estimator.
using a target for the power of a statistical test to be applied once the sample is collected.
using a confidence level, i.e. the larger the required confidence level, the larger the sample size (given a constant precision requirement).

UNIT - VI TEST OF HYPOTHESIS

Structure

6.0 INTRODUCTION

Hypothesis testing was introduced by Ronald Fisher, Jerzy Neyman, Karl Pearson and Pearson's son, Egon Pearson. Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter.

6.1 OBJECTIVES

- Understand purpose of Hypothesis testing;
- Understand the procedure for tests of hypotheses based on large samples
- Solve the problems of testing hypotheses concerning mean(s) and proportion(s) based on large samples

6.2 HYPOTHESIS TESTING ON POPULATION MEAN

Two situations may arise out of this, first one is when the population variance is known and the second situation is if the population variance is unknown.

6.2.1 POPULATION VARIANCE KNOWN

Steps:

1. Let **μ** and σ^2 be respectively the mean and the variance of the population under study, where σ^2 is known. If μ_0 is an admissible value of **μ**, then frame the null hypothesis as $H_0: \mu = \mu_0$ and choose the suitable alternative hypothesis from

   a. **(i) H1: $\mu \neq \mu0$ (ii) H1: $\mu > \mu0$ (iii) H1: $\mu < \mu0$**

2. Let (X1, X2, …,X_n) be a random sample of *n* observations drawn from the population, where **n** is large ($n \geq 30$).

3. Let the level of significance be ***α*.**

4. Consider the test statistics $Z= \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$. under H_0. Here $\bar{X}$ represents the sample mean, The approximate sampling distribution of the test statistics under H_0 is the N(0,1) distribution

5. Calculate the value of *Z* for the given sample (*x*1, *x*2, ..., *x*n) as

$$Z= \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}.$$

6. Find the critical value, *ze*, corresponding to *α* and *H*1 from the following table

Alternative Hypothesis (H_1)	$\mu \neq \mu0$	$\mu > \mu0$	$\mu < \mu0$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

7. Decide on *H0* choosing the suitable rejection rule from the following table corresponding to *H1*.

Alternative Hypothesis (H_1)	$\mu \neq \mu0$	$\mu > \mu0$	$\mu < \mu0$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example:

A company producing batteries finds that mean life span of the population of its batteries is 200 hours with a standard derivation of 15 hours. A sample of 100 batteries randomly chosen is found to have the mean life span of 195 hours. Test, at 5% level of significance, whether the mean life span of the batteries is significantly different from 200 hours.

Solution:

Step 1 : Let μ and σ represent respectively the mean and standard deviation of the probability distribution of the life span of the batteries. It is given that $\sigma = 15$ hours. The null and alternative hypotheses are

Null hypothesis: H_0: $\mu = 200$

i.e., the mean life span of the batteries is not significantly different from 200 hours.

Alternative hypothesis: $H_1 : \mu \neq 200$

i.e., the mean life span of the batteries is significantly different from 200 hours.

It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

Sample size (n) = 100, Sample mean ($\underline{x}$) = 195 hours

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic is , under H_0

Under the null hypothesis H_0, Z follows the $N(0,1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z under H_0 is calculated from

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}. = \frac{195-200}{15/\sqrt{100}}. = -3.33$$

Thus; $| z\ 0 | = 3.33$

Step 6 : Critical value

Since H_1 is a two-sided alternative, the critical value at $\alpha = 0.05$ is $z_e = z_{0.025} = 1.96$.

Step 7 : Decision

Since H_1 is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z_0| \geq z_e$. Thus, it is a two-tailed test. For the given sample information, the rejection rule holds *i.e.*, $|z_0| = 3.33 > z_e = 1.96$. Hence, H_0 is rejected in favour of $H_1: \mu \neq 200$. Thus, the mean life span of the batteries is significantly different from 200 hours.

6.2.2 POPULATION VARIANCE UNKNOWN

Steps:

1. Let μ and σ^2 be respectively the mean and the variance of the population under study, where σ^2 is unknown. If μ_0 is an admissible value of μ, then frame the null hypothesis as $H_0: \mu = \mu_0$ and choose the suitable alternative hypothesis from

   (i) $H_1: \mu \neq \mu_0$ (ii) $H_1: \mu > \mu_0$ (iii) $H_1: \mu < \mu_0$

2. Let $(X_1, X_2, \ldots, X_n)$ be a random sample of n observations drawn from the population, where n is large ($n \geq 30$).
3. Specify the level of significance, α.
4. Consider the test statistic $\mathbf{Z}= \frac{\bar{X}-\mu}{S/\sqrt{n}}$ under H_0, where X and S are the sample mean and sample standard deviation respectively. It may be noted that the above test statistic is obtained from Z by substituting S for σ.

   The approximate sampling distribution of the test statistic under H_0 is the N(0,1)distribution.

5. Calculate the value of Z for the given sample $(x_1, x_2, \ldots, x_n)$ as $\mathbf{Z}= \frac{\bar{X}-\mu}{S/\sqrt{n}}$. Here, $\bar{X}$ and s are respectively the values of $\bar{X}$ and S calculated for the given sample.
6. Find the critical value, z_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$\mu \neq \mu 0$	$\mu > \mu 0$	$\mu < \mu 0$

Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

7. Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu \neq \mu 0$	$\mu > \mu 0$	$\mu < \mu 0$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example:

A car manufacturing company desires to introduce a new model car . The company claims that the mean fuel consumption of its new model car is lower than that of the existing model of the car, which is 57 kms/litre. A sample of 100 cars of the new model car is selected randomly and their fuel consumptions are observed. It is found that the mean fuel consumption of the 100 new model car is 60 kms/litre with a standard deviation of 3 kms/litre. Test the claim of the company at 5% level of significance.

Solution:

Step 1 : Let the fuel consumption of the new model car be assumed to be distributed according to a distribution with mean and standard deviation respectively μ and σ. The null and alternative hypotheses are

Null hypothesis H_0: $\mu = 57$

i.e., the average fuel consumption of the company's new model car is not significantly different from that of the existing model.

Alternative hypothesis H_1: $\mu > 57$

i.e., the average fuel consumption of the company's new model car is significantly lower than that of the existing model. In other words, the number of kms by the new model csr is significantly more than that of the existing model car.

Step 2 : Data:

The given sample information are

Size of the sample (n) = 100. Hence, it is a large sample.

Sample mean ($\bar{x}$)= 30

Sample standard deviation(s) = 3

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under H_0 is

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

Since n is large, the sampling distribution of Z under H_0 is the $N(0,1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z for the given sample information is calculated from

$$Z_0 = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{60-57}{3/\sqrt{100}} = 10$$

Step 6 : Critical Value

Since H_1 is a one-sided (right) alternative hypothesis, the critical value at $\alpha = 0.05$ is $z_e = z_{0.05} = 1.645$.

Step 7 : Decision

Since H_1 is a one-sided (right) alternative, elements of the critical region are defined by the rejection rule $z_0 > z_e = z_{0.05}$. Thus, it is a right-tailed test. Since, for the given sample information, $z_0 = 10 > z_e = 1.645$, H_0 is rejected.

6.3 DIFFERENCE BETWEEN MEANS OF TWO POPULATIONS

6.3.1 POPULATION VARIANCE KNOWN

Steps:

1. Let μ_x and σ^2_x be respectively the mean and the variance of Population -1. Also, let μ_Y and σ^2_y be respectively the mean and the variance of Population -2 under study. Here σ^2_x and σ^2_y are known admissible values.

Frame the null hypothesis as H0: $\mu_X = \mu_Y$ and choose the suitable alternative hypothesis from

a. (i) H1: $\mu_X \neq \mu_Y$ (ii) H1: $\mu_X > \mu_Y$ (iii) H1: $\mu_X < \mu_Y$

2. Let (X1, X2, …, Xm) be a random sample of m observations

drawn from Population-1 and (Y1, Y2, …, Yn) be a random sample of n observations drawn from Population-2, where m and n are large(i.e., $m \geq 30$ and $n \geq 30$). Further, these two samples are assumed to be independent.

3. Specify the level of significance, α.

4. Consider the test statistic $Z = \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma^2_x}{m}+\frac{\sigma^2_y}{n}}}$ under H_0, where $\bar{X}$ and $\bar{Y}$ are respectively the means of the two samples described in Step-2.

The approximate sampling distribution of the test statistic $Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma^2_x}{m}+\frac{\sigma^2_y}{n}}}$ under H_0 (i.e., $\mu_X = \mu_Y$) is the N(0,1) distribution.

It may be noted that the test statistic, when $\sigma^2_x = \sigma^2_y = \sigma^2$, is $Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{1}{m}+\frac{1}{n}}}$

5. Calculate the value of Z for the given samples ($x_1, x_2, \ldots, x_m$) and ($y_1, y_2, \ldots, y_n$) as $Z0 = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{\sigma^2_x}{m}+\frac{\sigma^2_y}{n}}}$.

Here, $\bar{x}$ and $\bar{y}$are respectively the values of $\bar{X}$ and $\bar{Y}$for the given samples.

6. Find the critical value, *ze*, corresponding to *α* and *H1* from the following table

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

7. Make decision on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example:

Performance of students in a national level Olympiad exam was studied. The scores secured by randomly selected students from two districts, *viz*., D_1 and D_2 of a State were analyzed. The number of students randomly selected from D_1 and D_2 are respectively 1000 and 1600. Average scores secured by the students selected from D_1 and D_2 are respectively 116 and 114. Can the samples be regarded as drawn from the identical populations having common standard deviation 27 Test at 5% level of significance.

Solution:

Step 1 : Let μX and μY be respectively the mean scores secured in the national level Olympiad examination by all the students from the districts *D1* and *D2* considered for the study. It is given that the populations of the scores of the students of these districts have the common standard deviation σ = 2. The null and alternative hypotheses are

Null hypothesis: $H0: \mu_X = \mu_Y$

i.e., average scores secured by the students from the study districts are not significantly different.

Alternative hypothesis: $H1: \mu_X \neq \mu_Y$

i.e., average scores secured by the students from the study districts are significantly different. It is a two-sided alternative.

Step 2 : Data

The given sample information are

Size of the Sample-1 (m) = 1000

Size of the Sample-2 (n) = 1600. Hence, both the samples are large.

Mean of Sample-1 ($\bar{x}$) = 116

Mean of Sample-2 (y) = 114

Step 3 : Level of significance

α = 5%

Step 4 : Test statistic

The test statistic under the null hypothesis H_0 is $Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{1}{m}+\frac{1}{n}}}$

Since both *m* and *n* are large, the sampling distribution of *Z* under H_0 is the *N(0, 1)* distribution.

Step 5 : Calculation of Test Statistic

The value of *Z* is calculated for the given sample information from

$$Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{1}{m}+\frac{1}{n}}} = \frac{(116-114)}{\sqrt{\frac{1}{1000}+\frac{1}{1600}}} = 49.628$$

Step-6 : Critical value

Since H_1 is a two-sided alternative hypothesis, the critical value at $\alpha = 0.05$ is $ze = z0.025 = 1.96$.

Step-7 : Decision

Since H_1 is a two-sided alternative, elements of the critical region are defined by the rejection rule $|z_0| \geq z_e = z_{0.025}$. For the given sample information, $|z_0| = 49.628 > z_e = 1.96$. It indicates that the given sample contains sufficient evidence to reject H_0. Thus, it may be decided that H_0 is rejected. Therefore, the average performance of the students in the districts D_1 and D_2 in the national level Olympiad examination are significantly different. Thus the given samples are not drawn from identical populations.

6.3.2 POPULATION VARIANCE UNKNOWN

Steps:

1. Let μ_x and σ^2_x be respectively the mean and the variance of Population -1. Also, let μ_Y and σ^2_y be respectively the mean and the variance of Population -2 under study. Here σ^2_x and σ^2_y are known admissible values.

Frame the null hypothesis as H0: $\mu_X = \mu_Y$ and choose the suitable alternative hypothesis from

i. (i) H1: $\mu_X \neq \mu_Y$ (ii) H1: $\mu_X > \mu_Y$ (iii) H1: $\mu_X < \mu_Y$

2. Let (X1, X2, …, Xm) be a random sample of m observations drawn from Population-1 and (Y1, Y2, …, Yn) be a random sample of n observations drawn from Population-2, where m and

n are large(i.e., $m \geq 30$ and $n \geq 30$). Further, these two samples are assumed to be independent.

3. Specify the level of significance, α.

4. Consider the test statistic $Z = \frac{(\bar{X}-\bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{\frac{S^2_x}{m}+\frac{S^2_y}{n}}}$ under H_0,

5. (i.e., $\mu X = \mu Y$)

i.e, the above test stastistics is obtained from Z considered in the test described by substituting S^2_x and S^2_y respectively for σ^2_x and σ^2_y

The approximate sampling distribution of the test statistic $Z = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{S^2_x}{m}+\frac{S^2_y}{n}}}$ under H_0 is the N(0,1) distribution.

6. Calculate the value of Z for the given samples ($x_1, x_2, ..., x_m$) and ($y_1, y_2, ..., y_n$) as $Z0 = \frac{(\bar{X}-\bar{Y})}{\sqrt{\frac{S^2_x}{m}+\frac{S^2_y}{n}}}$.

Here, $\bar{x}$ and $\bar{y}$are respectively the values of $\bar{X}$ and $\bar{Y}$for the given samples.

Also, s^2_x and $s^2_y$are respectively the values of S^2_x and S^2_y for the given samples.

7. Find the critical value, z_e, corresponding to α and $H1$ from the following table

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

8. Make decision on H_0 choosing the suitable rejection rule from the following table corresponding to H_1.

Alternative Hypothesis (H_1)	$\mu_X \neq \mu_Y$	$\mu_X > \mu_Y$	$\mu_X < \mu_Y$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

6.4 TEST OF HYPOTHESES FOR POPULATION PROPORTION

Steps:

1. Let P denote the proportion of the population possessing the qualitative characteristic (attribute) under study. If p_0 is an admissible value of P, then frame the null hypothesis as $H_0: P = p_0$ and choose the suitable alternative hypothesis from

   (i) $H_1: P \neq p_0$ (ii) $H_1: P > p_0$ (iii) $H_1: P < p_0$

2. Let p be proportion of the sample observations possessing the attribute, where n is large, $np > 5$ and $n(1 - p) > 5$.

3. Specify the level of significance, α.
4. Consider the test statistic under H_0. Here, $\frac{p-P}{\sqrt{\frac{PQ}{n}}}$, Here, $Q = 1 - P$.

   The approximate sampling distribution of the test statistic under $H0$ is the $N(0,1)$ distribution.

5. Calculate the value of Z under H_0 for the given data as , $\frac{p-P}{\sqrt{\frac{p0q0}{n}}}$, $q_0 = 1 - p_0$.
6. Choose the critical value, ze, corresponding to α and $H1$ from the following table

Alternative Hypothesis (H_1)	$P \neq p_0$	$P > p_0$	$P < p_0$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

7. Make decision on $H0$ choosing the suitable rejection rule from the following table corresponding to $H1$.

Alternative Hypothesis (H_1)	$P \neq p_0$	$P > p_0$	$P < p_0$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example:

A survey was conducted among the students of a city to study their preference towards consumption of chocolate and ice-cream. Among 2000 randomly selected students, it is found that 1120 are chocolate and the remaining are ice-cream. Can we conclude at 1% level of significance from this information that both chocolate and ice-cream are equally preferred among the students in the city?

Solution:

Step 1 : Let P denote the proportion of students in the city who preferred to have chocolate. Then, the null and the alternative hypotheses are

Null hypothesis: H_0 : = 0.5

i.e., it is significant that both chocolate and ice-cream are preferred equally in the city.

Alternative hypothesis: H_0 : ≠ 0.5

i.e., preference of chocolate and ice-cream are not significantly equal. It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

Sample size (n) = 2000. Hence, it is a large sample.

No. of chocolate consumer = 1120

Sample proportion (p) $=\frac{1120}{2000} = 0.56$

Step 3 : Level of significance

$\alpha = 1\%$

Step 4 : Test statistic

Since n is large, np= 1120 > 5 and $n(1 - p)$ = 880 > 5, the test statistic under the null hypothesis, is $Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}}$

Its sampling distribution under $H0$ is the $N(0,1)$ distribution.

Step 5 : Calculation of Test Statistic

The value of Z can be calculated for the sample information from

$$Z = \frac{p-P}{\sqrt{\frac{PQ}{n}}} = \frac{0.56-0.50}{\sqrt{\frac{0.5 \, x \, 0.5}{2000}}} = 5.3763$$

Step 6 : Critical value

Since $H1$ is a two-sided alternative hypothesis, the critical value at 1% level of significance is $z_{\alpha/2} = z_{0.005} = 2.58$.

Step 7 : Decision

Since $H1$ is a two-sided alternative, elements of the critical region are determined by the rejection rule $|z0| \geq ze$. Thus it

is a two-tailed test. Since $|z0| = 5.3763 > ze = 2.58$, reject $H0$ at 1% level of significance. Therefore, there is significant evidence to conclude that the preference of chocolate and ice-cream are different.

6.5 DIFFERENCE BETWEEN TWO PROPORTIONS

Steps:

1: Let P_X and P_Y denote respectively the proportions of Population-1 and Population-2 possessing the qualitative characteristic (attribute) under study. Frame the null hypothesis as H_0: PX=PY and choose the suitable alternative hypothesis from

(i) $H_1: P_X \neq P_Y$ (ii) $H_1: P_X > P_Y$ (iii) $H_1: P_X < P_Y$

2: Let P_X and P_Y denote respectively the proportions of the samples of sizes m and n drawn from Population-1 and Population-2 possessing the attribute, where m and n are large (i.e., $m \geq 30$ and $n \geq 30$). Also, $mp_x > 5$, $m(1- p_x) > 5$, $np_y > 5$, $n(1- p_y) > 5$. Here, these two samples are assumed to be independent.

3: Specify the level of significance, α.

4: Consider the test statistic $Z = \frac{(px-py)-(PX-PY)}{\sqrt{\bar{p}\bar{q}(\frac{1}{m}+\frac{1}{n})}}$ under H_0. Here

$\hat{p} = \frac{mpx + npy}{m+n}$, $\hat{q} = 1 - \hat{p}$. The approximate sampling distribution of the test statistic

under H_0 is N(0,1) distribution.

5: Calculate the value of Z for the given data as $Z = \frac{(PX-PY)}{\sqrt{\bar{p}\bar{q}(\frac{1}{m}+\frac{1}{n})}}$

6: Choose the critical value, z_e, corresponding to α and H_1 from the following table

Alternative Hypothesis (H_1)	$P_X \neq P_Y$	$P_X > P_Y$	$P_X < P_Y$
Critical Value (z_e)	$z_{\alpha/2}$	z_α	$-z_\alpha$

7: Decide on H_0 choosing the suitable rejection rule from the following table corresponding to H_1

Alternative Hypothesis (H_1)	$P_X \neq P_Y$	$P_X > P_Y$	$P_X < P_Y$
Rejection Rule	$\lvert z_0 \rvert \geq z_{\alpha/2}$	$z_0 > z_\alpha$	$z_0 < -z_\alpha$

Example:

A study was conducted to investigate the interest of students in private schools. Among randomly selected 1000 students from City-1, 800 persons were found to be private school. From City-2 , 1600 persons were selected randomly and among them 1200 students are from private school. Do the data indicate that the two cities are significantly different with respect to prevalence of private school among the students? Choose the level of significance as $\alpha = 0.05$.

Solution:

Step1 : Let *PX* and *PY* be respectively the proportions of private school students in City-1 and City-2. Then, the null and alternative hypotheses are

Null hypothesis: H_0: $P_X = P_Y$

i.e., there is no significant difference between the proportions of private school students in City-1 and City-2.

Alternative hypothesis: H_1: $P_X \neq P_Y$

i.e., difference between the proportions of private school students in City-1 and City-2 is significant. It is a two-sided alternative hypothesis.

Step 2 : Data

The given sample information are

City	Sample size	Sample proportion
City 1	m = 1000	$P_X = 800/1000 = 0.80$
City 2	n = 1600	$P_Y = 1200/1600 = 0.75$

Here $m \geq 30$ and $n \geq 30$, $mp_x = 800 > 5$,

$m(1 - p_x) = 200 > 5$, $np_y = 1200 > 5$, $n(1 - p_y) = 400 > 5$.

Step 3 : Level of significance

$\alpha = 5\%$

Step 4 : Test statistic

The test statistic under the null hypothesis is

$$Z = \frac{(px - py) - (PX - PY)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{m}+\frac{1}{n}\right)}} \quad \text{where } \hat{p} = \frac{mpx + npy}{m+n},\ \hat{q} = 1 - \hat{p}$$

Step 5 : Calculation of Test Statistic

The value of *Z* for given sample information is calculated from

$$Z = \frac{(PX - PY)}{\sqrt{\bar{p}\bar{q}\left(\frac{1}{m}+\frac{1}{n}\right)}} = \frac{(0.80 - 0.75)}{\sqrt{(0.77)(0.23)\left(\frac{1}{1000}+\frac{1}{1600}\right)}} = 2.0764$$

Step 6 : Critical value

Since *H1* is a two-sided alternative hypothesis, the critical value at 5% level of significance is *ze*= 1.96.

Step 7 : Decision

Since *H0* is a two-sided alternative, elements of the critical region are determined by the rejection rule |z0| >*ze*. Thus, it is a two-tailed test. For the given sample information, z_e = 2.0764 > = 1.96. Hence, *H0* is rejected. We can conclude that the difference between the proportions of private school students in City-1 and City-2 is significant.

CHECK YOUR PROGRESS - 1

1. What is hypothesis testing?
2. A large sample theory is applicable when
3. When is standard error of the sample proportion under H_0

6.6 SUMMARY

- If the number of sample observations is greater than or equal to 30, it is called large sample.
- For hypotheses testing based on two samples, if the sizes of both the samples are greater than or equal to 30, they are called large samples.
- For testing population proportion, the sampling distribution of the test statistic is *N(0, 1)*, only when $n \geq 30$, $np > 5$ and $n(1 - p) > 5$.
- For testing equality of two population proportions, the sampling distribution of the test statistic is *N (0, 1)* only when $m \geq 30$, $n \geq 30$, $mpX > 5$, $m(1 - pX) > 5$, $npY > 5$ and $n(1 - pY) > 5$.

6.7 KEY WORDS

6.8 ANSWER TO CHECK YOUR PROGRESS

1. Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data
2. When $n \geq 30$
3. $\sqrt{\frac{PQ}{n}}$

6.9 QUESTION AND EXERCISE

SHORT ANSWER QUESTIONS

1. List the possible alternative hypotheses and the corresponding rejection rules followed in testing equality of two population means.
2. Specify the alternative hypotheses and the rejection rules prescribed for testing equality of two population proportions

LONG ANSWER QUESTIONS

1. Explain the general procedure to be followed for testing of hypotheses.
2. Explain the procedure for testing hypotheses for population mean, when the population variance is unknown.
3. How will you formulate the hypotheses for testing equality of means of two populations, when the population variances are known? Describe the method.
4. Describe the procedure for testing hypotheses concerning equality of means of two populations, assuming that the population variances are unknown.
5. Give a detailed account on testing hypotheses for population proportion.
6. Explain the procedure of testing hypotheses for equality of proportion of two populations. Interest of XII Students on Residential Schooling was investigated among randomly selected

students from two regions. Among 300 students selected from Region A, 34 students expressed their interest. Among 200 students selected from Region B, 28 students expressed their interest. Does this information provide sufficient evidence to conclude at 5% level of significance that students in Region A are more interested in Residential Schooling than the students in Region B?

6.10 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawanPublishersand Distributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills PublishingCompany Ltd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw HillPublishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd.,New Delhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons.,NewDelhi.
6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.
7. Statistics for Management by Richard I. Levin and David S. Rubin. PrenticeHall of India Pvt. Ltd., New Delhi.

UNIT 7 - CHI – SQUARE TEST

Structure

7.0 INTRODUCTION

A chi-squared test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test.

The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

Chi square test is applied in statistics to test the goodness of fit to verify the distribution of observed data with assumed theoretical distribution. Therefore, it is a measure to study the divergence of actual and excepted frequencies. it has great use in statistics, specially in sampling studies, where we except a doubted coincidence between actual and excepted frequencies, and the extent to which the difference can be ignored, because of fluctuations in sampling.

7.1 OBJECTIVES

The student will be able to

- Understand the purpose for using chi-square test
- Understand the procedures for Analysis of variance
- Understand the characteristics and of chi-square test
- Solve problems to test the hypothesis whether the population has a particular variance using chi-square test

7.2 CHARACTERISTICS OF χ^2 TEST

1. Test is based on events or frequencies, whereas in theoretical distribution, the test is based on mean and standard deviation.
2. To draw inferences, this test is applied, specially testing the hypothesis but not useful for estimation.
3. The test can be used between the entire set of observed and excepted frequencies.
4. For every increase in the number of degree of freedom, a new $\chi 2$ distribution is formed.
5. It is a general purpose test and as such is highly useful n research.

7.3 USES OF χ^2 TEST

χ^2 Test of goodness of fit

Through the test we can find out the deviations between the observed values and excepted values. Here we are not concerned with the parameters but concerned with the form of distribution. Karl Pearson has developed a method to test the difference between the theoretical value (hypothesis) and the observed value. A Greek letter χ^2 is used to describe the magnitude of difference between the fact and theory.

The χ^2may be defined as,

$$\chi^2 = \frac{O - E^2}{E}$$

O = Observed Frequencies

E = Excepted Frequencies

7.4 STEPS OF χ^2 TEST

1. A hypothesis is established along with the significance level.
2. Compute deviation between observed value and excepted value (O-E).
3. Square the deviations calculated $(O\text{-}E)^2$.
4. Divide the $(O\text{-}E)^2$ by its excepted frequency.
5. Add all the values obtained in step 4.
6. Find the value of χ^2 table at certain level of significance, usually 5% level.

If the calculated value of χ^2 is greater than the table value of χ^2, at certain level of significance, we reject the hypothesis. If the computed value of $\chi^{2\ is}$ less than the table value, at a certain degree of level of significance, it is said to be non-significant. This implies that the discrepancy between the observed frequencies may be due to fluctuations in the simple sampling.

Example:

In a certain ample of 2000 families, 1400 families are consumers of tea. Out of 1800 Hindu families, 1236 families consume tea. Use χ^2test and state whether there is any significant difference between consumption of tea among Hindu and non-Hindu families.

	Hindu	Non – Hindu	Total
Consuming Tea	1236	164	1400
Non – Consuming Tea	564	36	600
Total	1800	200	2000

Solution

On tabulation of the information in a 2x2 contingency table, we get:

Observed Frequencies

	Hindu	**Non – Hindu**	**Total**
Consuming Tea	1236	164	1400
Non – Consuming Tea	564	36	600
Total	1800	200	2000

Excepted Frequencies

	Hindu	**Non – Hindu**	**Total**
Consuming Tea	1260	140	1400
Non – Consuming Tea	540	60	600
Total	1800	200	2000

Calculation of χ^2

O	**E**	**O – E**	**$(O-E)^2$**	**$(O-E)^2/E$**
1236	1260	-24	576	0.457
564	540	24	576	1.068
164	140	24-24	576	4.114
36	60		576	9.600
				$\sum(O-E)^2/E=15.239$

d.f is 1, Table value of $r2$ 0.05 for 1 d.f = 3.841.
For a contingency table, 2x2 table, the degree of freedom is

$V = (c-1)(r-1) = (2-1)(2-1) = 1.$

The calculated value of χ^2 15.239 is higher than the table value i.e., 3.841; therefore the null hypothesis is rejected.

Hence, the two communities differ significantly as far as consumption of a tea is concerned.

7.5. SUMMARY

- The uses of distribution are testing the specified variance of a normal population, testing goodness of fit and testing independence of attributes
- Through the test we can find out the deviations between the observed values and excepted values. Here we are not concerned with the parameters but concerned with the form of distribution.

7.6 Questions & Exercises

SHORT ANSWER QUESTION:

1. Define chi square test
2. What are the condition for the validity of chi square test
3. Five applications of chi square test
4. What are the uses of chi square

LONG ANSWER QUESTION:

1. Explain the steps of chi-square test
2. Write down steps for testing the significance of goodness of fit

7.11 FURTHER READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics., London
2. McGraw Hill Book Company.
3. Yamane, T.: Statiscs: An Introductory Analysis, New York, HarperedRow Publication
4. R.P. Hooda: Statistic for Economic and Management McMillan IndiaLtd.
5. G.C. Beri: Statistics for Mgt., TMA
6. J.K. Sharma: Business Statistics, Pearson Education

Unit – VIII F – TEST

Structure

8.1 Introduction

An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis . It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.

8.2 ANALYSIS OF VARIANCE (ANOVA)

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures (such as the "variation" among and between groups) used to analyze the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether two or more population means are equal, and therefore generalizes the t-test beyond two means.

8.3 ASSUMPTIONS IN ANALYSIS OF VARIANCE

1. Each of the samples is strained from a normal distribution.
2. The variances for the population from which samples have been drained are equal.
3. The variation of each value around its own grand mean should be independent for each value.

8.4 BASIC STEPS IN ANALYSIS OF VARIANCE

Determine

1. One estimate of the population variance from the variance among the sample means.
2. Determine a second estimate of the population variance from the variance within the sample.
3. Compare these two estimates if they are approximately equal in value, accept the null hypothesis.

8.4.1 One-Way Anova

In statistics, one-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare means of two or more samples (using the F distribution). This technique can be used only for numerical response data, the "Y", usually one variable, and numerical or (usually) categorical input data, the "X", always one variable, hence "one-way".

The ANOVA tests the null hypothesis that samples in all groups are drawn from populations with the same mean values. To do this, two estimates are made of the population variance. These estimates rely on various assumptions.

The ANOVA produces an F-statistic, the ratio of the variance calculated among the means to the variance within the samples.When there are only two means to compare, the t-test and the F-test are equivalent; the relation between ANOVA and t is given by $F = t^2$. An extension of one-way ANOVA is two-way analysis of variance that examines the influence of two different categorical independent variables on one dependent variable.

Example:

In order to determine whether there is significant difference in the durability of 3 makes of computers, samples of size 5 are selected from each make and the frequency of repair during the 1st year of purchase is observed. The results are as follows:

Makes		
I	II	III
4	7	6
6	9	4
8	11	6
9	12	3
7	5	2

In view of the above data, what conclusion can you draw?

Solution:

Null Hypothesis H_o= there is no significant difference in the durability of 3 makes of computers.

Computer I		**Computer II**		**Computer III**	
X_1	X_1^2	X_2	X_2^2	X_3	X_3^2
4	16	7	49	6	36
6	36	9	81	4	16
8	64	11	121	6	36

9	81	12	144	3	9
7	49	5	25	2	4
$\sum X_1$=34	$\sum X_1^2$=246	$\sum X_2$=44	$\sum X_2^2$=420	$\sum X_3$=21	$\sum X_3^2$=101

Step – 1

Sum of all items (T) = $\sum X_1+\sum X_2+\sum X_3$
= 34+44+21
=99

Step – 2

Correction factor (C.F) = $\frac{T^2}{N} = \frac{(99)^2}{15} =$ 653.4

Step – 3

TSS = Sum of Squares of all the items – C.F

$= \sum X_1^2+\sum X_2^2+\sum X_3^2 - \frac{T^2}{N}$

= 246+420+101-653.4= 113.6

Step – 4

SSC = Sum of Squares between samples – C.F

$= \frac{(\sum X_1)^2}{n} + \frac{(\sum X_2)^2}{n} + \frac{(\sum X_3)^2}{n} -$ C.F

$= \frac{(34)^2}{5} + \frac{(44)^2}{5} + \frac{(21)^2}{5} -$ 653.4

= 231.2 + 387.2 + 88.5 – 653.= 53.5

Step – 5

MSC = $\frac{\text{Sum of squares between samples}}{\text{d.f}}$

$= \frac{53.5}{2}$

= 26.75

Step – 6

SSE = Total sum of squares – Sum of Squares between samples

= 113.6 – 53.5
= 60.1

Step – 7

MSE = $\frac{\text{Sum of squares within samples}}{\text{d.f}}$

$= \frac{60.1}{12}$

= 5.00

ANOVA TABLE

Source of variations	**Sum of squares**	**Degrees of freedom**	**Mean Squares**	**F - ratio**
Between samples	SSC = 53.5	3-1=2	MSC= $\frac{SSC}{d.f}$	

			= 26.75	$F_c = \frac{MSC}{MSE}$ = 5.35
Within samples	SSE = 60.1	15-3=12	MSE= $\frac{SSE}{d.f}$ = 5.00	

Tabulated value of F for V_1=2 and V_2=12 at 5% level of significance is 3.88. F_{Tab}=3.88. Calculated value of F is F_c = 5.35. Since $F_c > F_{Tab.}$ We reject the null hypothesis $H_{0.}$ There is significant difference in the durability of 3 makes of computers.

8.4.2 TWO-WAY ANOVA

The two-way ANOVA compares the mean differences between groups that have been split on two independent variables (called factors). The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable.

This is a term which stems from agricultural research in which several variables or treatments are applied to different blocks of land for repetition or replication of the experimental effects. The advantages of a completely randomized experimental design are as follows.

(a) Easy to lay out.
(b) Allows flexibility.
(c) Simple statistical analysis.
(d) The lot of information due to missing data is smaller than with any other design.

But this design is usually suited (i) only for small number of treatments and (ii) for homogeneous experimental material.

Example:

There varieties A, B, C of crop are tested in a randomized block design with four replications. The plot yields in pounds are as follows

Varieties	Yields			
	1	2	3	4
A	6	5	8	9
B	8	4	6	9
C	7	6	10	6

Solution

Null hypothesis H_0: There is no significant difference between varieties (rows) and between yields,(blocks).

Varieties	Yields				
	1	2	3	4	Total
A	6	4	6	6	24
B	7	5	8	9	28
C	8	6	10	9	32
Total	21	15	24	24	84

Step -1

Grand total (T) = 84

Step – 2

Correction factor (C.F) = $\frac{T2}{N} = \frac{(84)2}{12}$ = 588

Step – 3

SSC = Sum of squares between blocks (columns)

$= \frac{(21)^2}{3} + \frac{(15)^2}{3} + \frac{(24)^2}{3} + \frac{(24)^2}{3} - C.F$

= 606 – 588

= 18

Step – 4

SSR = Sum of squares between varieties (Rows)

$= \frac{(24)^2}{4} + \frac{(28)^2}{4} + \frac{(32)^2}{4} - C.F$

= 596 – 588

= 8

Step – 5

TSS = Total sum of squares – C.F

$= [(6)^2+(7)^2+(8)^2+(4)^2+(6)^2+(5)^2+(8)^2+(6)^2+(10)^2+(6)^2+(9)^2+(9)^2] - 588$

= 624 – 588

= 36

Step – 6

SSE = Residual sum of squares

= TSS-(SSC+SSR)

= 36 – (18+8) = 10

Step- 7

d.f = v3 = (c-1) (r-1)

= (3) (2)

= 6

ANOVA TABLE

Source of variation	**Sum of squares**	**Degree of freedom**	**Mean Squares**	**F-ratio**
Between Blocks (Columns)	SSC= 18	c-1 4-1= 3	MSC= $\frac{SSC}{d.f}$ = 6	$F_c = \frac{MSC}{MSE}$ = 3.6
Between Varieties (Rows)	SSR=8	r-1 3-1=2	MSR= $\frac{SSR}{d.f}$ = 4	$F_R = \frac{MSR}{MSE}$ = 2.4
Residual	SSE=10	(r-1)(c-1) = 6	MSE= $\frac{SSE}{d.f}$ =1.667	

(i) The tabulated value of F for (3,6) d.f at 5 % level of significance is 4.76.F_{tab}=4.76. since $F_c < F_{tab}$, we accept the null hypothesis H_0. That is there is no significant difference between yields.

(ii) The tabulated value of F for (2,6) d.f at 5 % level of significance is 5.14.F_{tab}=5.14. since $F_R < F_{tab}$, we accept the null hypothesis H_0. That is there is no significant difference between varieties.

CHECK YOUR PROGRESS - 1

1. By which other name is the Chi-Square goodness of fit test known?
2. What type of data do you need in Chi-Square test?
3. What symbol is used to represent Chi-Square?
4. What is Analysis of Variance?
5. What is the main purpose of Two way ANOVA test?
6. The variation of each value around its own grand mean should be _______________ for each value

8.5 SUMMARY

- The uses of distribution are testing the specified variance of a normal population, testing goodness of fit and testing independence of attributes
- Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures used to analyze the differences among group means in a sample
- One-way analysis of variance (abbreviated one-way ANOVA) is a technique that can be used to compare means of two or more samples
- The two-way ANOVA compares the mean differences between groups that have been split on two independent variables

8.6 KEY WORDS

Chi-square, Analysis of Variance, One way method, Two way method

8.7 ANSWER TO CHECK YOUR PROGRESS

1. One sample chi square
2. Categorical
3. $\chi 2$
4. ANOVA provides a statistical testof whether two or more population means are equal, and therefore generalizes the *t*-test beyond two means
5. The primary purpose of a two-way ANOVA is to understand if there is an interaction between the two independent variables on the dependent variable
6. Independent

8.8 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTION:

1. Define chi square test
2. What are the condition for the validity of chi squate test
3. Five applications of chi square test
4. What is analysis of variance?
5. What are the assumptions of ANOVA

LONG ANSWER QUESTION:

1. Explain the steps of chi-square test
2. Write down steps for testing the significance of goodness of fit
3. Write the model ANOVA table for one way classification
4. Compare one way and two way ANOVA

8.9 FURTHER READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics., London
2. McGraw Hill Book Company.
3. Yamane, T.: Statiscs: An Introductory Analysis, New York, HarperedRow Publication
4. R.P. Hooda: Statistic for Economic and Management McMillan IndiaLtd.
5. G.C. Beri: Statistics for Mgt., TMA
6. J.K. Sharma: Business Statistics, Pearson Education

UNIT IX - CORRELATION ANALYSES

Structure

9.0 INTRODUCTION

In our day to day life, we find many situations when a mutual relationship exists between two variables i.e., with change (fall or rise) in the value of one variable there may be change (fall or rise) in the value of other variable For example, as price of a commodity increases the demand for the commodity decreases. In the increase in the levels of pressure, the volume of a gas decreases at a constant temperature. These facts indicate that there is certainly some mutual relationships that exist between the demand of a commodity and its price, and pressure and volume. Such association is studied in correlation analysis. The correlation is a statistical tool which measures the degree or intensity or extent of relationship between two variables and correlation analysis involves various methods and techniques used for studying and measuring the extent of the relationship between the two variables.

9.1 OBJECTIVES

After studying this chapter students will be able to understand

- Understand the concept of scatter Diagram
- Concept of Karl Pearson's correlation co-efficient and the

methods of computing it.

- Spearman's Rank correlation co-efficient

9.2 CORRELATION

Correlation is a statistical technique which measures and analyses the degree or extent to which two or more variables fluctuate with reference to one another. It denotes the inter-dependence amongst variables. The degrees are expressed by a coefficient which ranges between -1 to +1. The direction of change is indicated by + or - signs; the former, refers to the movement in the same direction and the later, in the opposite direction. An absence of correlation is indicated by zero. Correlation thus expresses the relationship through a relative measure of change and it has nothing to do with the units in which the variables are expressed.

9.3 LINEAR CORRELATION

If the amount of change in one variable tends to bear constant ratio to the amount of change in the other variable then the correlation is said to be linear. For example,

X	5	10	15	20	25
Y	90	170	230	310	420

9.4 TYPES OF CORRELATION

There are three important types of correlation. They are

1. Positive and Negative correlation
2. Simple, Partial and Multiple correlation
3. Linear and Non-Linear correlation

1. Positive and Negative correlation

Correlation is classified according to the direction of change in the two variables. In this regard, the correlation may either be positive or negative.

Positive correlation refers to the change (movement)of variables in the same direction. Both the variables are increased or decreased in the same direction, it is called positive correlation. It is otherwise called as direct correlation. For example, a positive correlation exists between ages of husband and wife, height and weight of a group of individuals, increase in rainfall and production of paddy, increase in the offer and sales.

Negative correlation refers to the change (movement) of variables in the opposite direction. In other words, an increase (decrease) in the value of one variable is followed by a decrease by a decrease (increase) in the value of the other is said to be negative correlation. It is otherwise called increase correlation. For example, a negative correlation exists

between price and demand, yield of crop and price.

The following expels illustrate the concept of positive correlation and negative correlation.

Positive correlation

X	5	7	9	11	16	20	28
y	20	26	35	37	48	50	55

Negative Correlation

X	14	17	23	35	46
y	16	12	10	9	5

2.Simple, Partial and Multiple Correlations

Simple correlation is a measure used to determine the strength and the direction of the relationship between two variables, X and Y. A simple correlation coefficient can range from –1 to 1. However, maximum (or minimum) values of some simple correlations cannot reach unity (i.e., 1 or –1).

When we study only two variables, the relationship is described as simple correlation; example, quantity of money and price level, demand and price, etc. But in a multiple correlation we study more than two variables simultaneously; example, the relationship of price, demand and supply of a commodity.

The study of two variables excluding some other variables is called partial correlation. For example, we study price and demand, eliminating the supply side.

3. Linear and Non-Linear Correlation

Linear correlation is a measure of the degree to which two variables vary together, or ameasure of the intensity of the association between two variables.

If the ratio of change between two variables is uniform, then the there will be linear correlation between them. Consider the following.

X	6	12	18	24
Y	5	10	15	20

The ratio of change between the variables is same.

In a curvilinear or non linear correlation, the amount of change in one variable does not bear a constant ratio of the amount of change in the

other variables. The graph of non-linear or curvilinear relationship will form a curve.

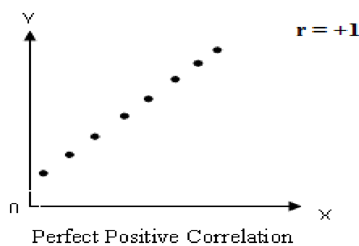
In majority of cases, we find curvilinear relationship, which is a complicated one, so we generally assume that the relationship between the variables under the study is linear. In social sciences, linear correlation is rare, because the exactness is not as perfect as in natural sciences.

CHECK YOUR PROGRESS- 1

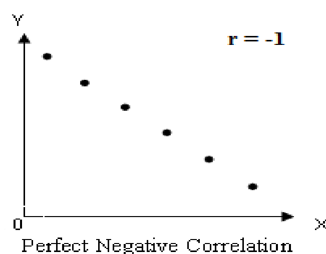
1. What is correlation?
2. Define linear correlation?
3. List out the different types of correlation?

9.5 SCATTER DIAGRAM

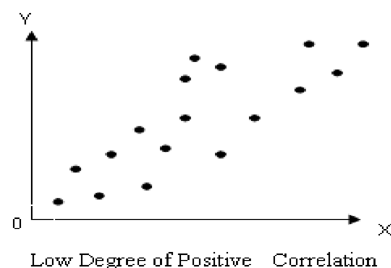
It is simple and attractive method of diagrammatic representation. In this method, the given data are plotted on a graph sheet in the form of dots. The x variables are plotted on the horizontal axis and y variables on the vertical axis. Now we can know the scatter or concentration of the various points. This will show the type of correlation.



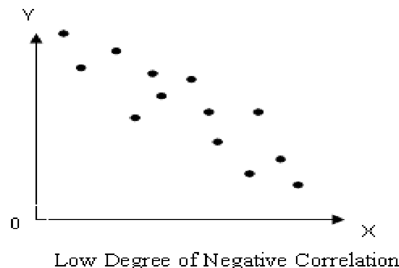
Perfect Positive Correlation



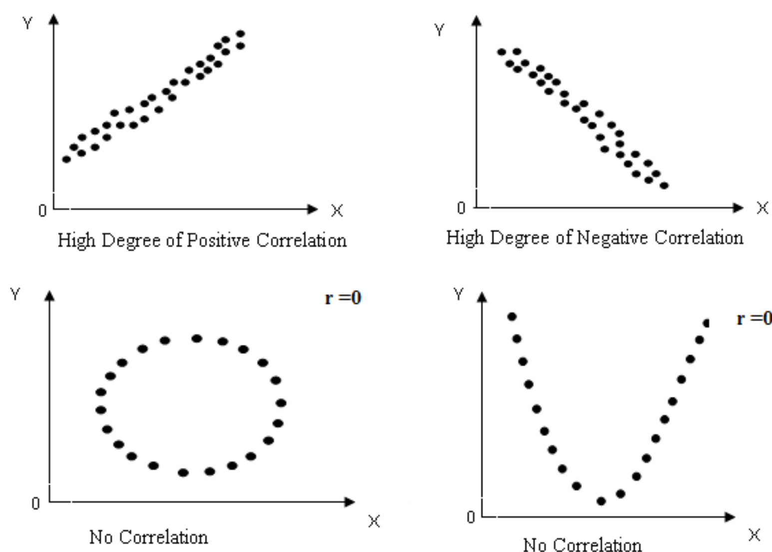
Perfect Negative Correlation



Low Degree of Positive Correlation



Low Degree of Negative Correlation



9.6 TWO-WAY TABLE

Atwo-way table (also called a contingency table) is a useful tool for examining relationships between categorical variables; the entries in the cells of two-way table can be frequency counts or relative frequencies (just like a one-way table).

	Dance	**Sports**	**TV**	**Total**
Men	2	10	8	20
Women	16	6	8	30
Total	18	16	16	50

Above a two-way table shows the favourite leisure activities for 50 adults-20 men and 30 women. Because entries in the table are frequency counts, the table is a frequency table.

9.7 PEARSON'S CO-EFFICIENT OF CORRELATION

Karl Pearson (1867-1936), the British biometrician suggested this method. It is popularly known as Pearson's co-efficient of correlation. It is mathematical method for measuring the magnitude of linear relationship between two variables.

Pearson's correlation coefficient is the test statistics that measures the statistical relationship, or association, between two continuous variables. It is known as the best method of measuring the association between variables of interest because it is based on the method of covariance.

(a)Arithmetic mean Method

$$r=\frac{\sum xy}{\sqrt{\sum x2 \sum y2}}$$

Example:

Find Pearson's Co-efficient of correlation from the following data

Sales	15	18	22	28	32	46	52
Profit	52	66	78	87	96	125	141

Solution

Let the sales be denoted by x and the profit by y.

Computation of coefficients of correlation

X	$\mathbf{X-\bar{X}}$	$\mathbf{X^2}$	**Y**	$\mathbf{Y-\bar{Y}}$	$\mathbf{Y^2}$	**XY**
15	-15.43	238.98	52	-40.14	1611.22	619.36
18	-12.43	154.50	66	-26.14	683.30	324.92
22	-8.43	71.06	78	-14.14	199.94	119.20
28	-2.43	5.90	87	-5.14	26.42	12.49
32	1.57	2.46	96	3.86	14.90	6.06
46	15.57	242.42	125	32.86	1079.78	511.63
52	21.57	465.26	141	48.86	2387.30	1053.91
$\sum$**x =213**	$\sum$**x=-0.01**	$\sum$**x^2=1179.68**	$\sum$**y=645**	$\sum$**y= 0.02**	$\sum$**y^2 = 6,002.86**	$\sum$**xy =2647.57**

X=∑x/N =213/7=30.43

Y=∑y/N =645/7 =92.14

$$\sum x^2=1179.68, \sum y^2=6002.86, \sum xy=2647.57$$

$$r=\frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}=\frac{2647.57}{\sqrt{1179.68x6,002.86}}=\frac{2647.57}{\sqrt{1179.68x6,002.86}}$$

$$=\frac{2647.57}{34.35x77.48}=\frac{2647.57}{2661.44}=0.99$$

Therefore, there is a high degree positive correlation between the x and y.

9.8 SPEARMEN'S RANK CORRELATION CO-EFFICIENT

In statistics, Spearman's rank correlation coefficient or Spearman's rho, named after Charles Spearman and often denoted by the Greek letter P(rho)or as r_s is a nonparametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function.

The Spearman correlation between two variables is equal to the Pearson correlation between the rank values of those two variables; while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not). If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$$

Spearman's coefficient is appropriate for both continuous and discrete ordinal variables. Both Spearmen's can be formulated as special cases of a more general correlation coefficient.

Example:

Two faculty members ranked 12 candidates for scholarships. Calculate the spearman rank correlation coefficient.

Candidate	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**

Professor A	**8**	**12**	**6**	**4**	**9**	**15**	**8**	**7**	**16**	**13**
Professor B	**9**	**16**	**10**	**8**	**14**	**19**	**12**	**11**	**20**	**17**

Solution

R_x	R_y	$d= R_x- R_y$	d^2
8	9	-1	1
12	16	-4	16
6	10	-4	16
4	8	-4	16
9	5	4	16
15	10	5	25
8	7	1	1
7	11	-4	16
16	15	1	1
13	18	-5	25
			$\sum d^2=133$

$$r_s = 1 - \frac{6\sum D^2}{n(n^2-1)} = 1 - \frac{6(133)}{10(100-1)} = 1 - \frac{798}{990}$$

= 1-0.8060

r = 0.194

CHECK YOUR PROGRESS - 2

4. What are the uses of scattered diagram?
5. Write the formula to calculate spearman's rank correlation?

9.9 PROPERTIES OF CORRELATION CO-EFFICIENT

1. Coefficient of Correlation lies between -1 and +1:The coefficient of correlation cannot take value less than -1 or more than one +1. Symbolically,-1<=r<= + 1 or | r | <1.

2. Coefficients of Correlation are independent of Change of Origin: This property reveals that if we subtract any constant from all the values of X and Y, it will not affect the coefficient of correlation.
3. Coefficients of Correlation possess the property of symmetry: The degree of relationship between two variables is symmetric.
4. **Coefficient of Correlation is independent of Change of Scale:** This property reveals that if we divide or multiply all the values of X and Y, it will not affect the coefficient of correlation.
5. The value of the co efficient of correlation shall always lie between +1 and -1.
6. When r = + 1, then there is perfect positive correlation between the variables.
7. When r = - 1, then there is perfect negative correlation between the variables.
8. When r = 0, then there is no relationship between the variables.

The third formula given above, that is

$$r=\frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

It is easy to calculate, and it is not necessary to calculate the standard deviation of X and Y series separately.

9.10 SUMMARY

- The term correlation refers to the degree of relationship between two or more variables.
- Scatter diagram is a graphic device for finding correlation between two variables.
- Karl Pearson correlation coefficient $r(x,y)$ = $r=\frac{\sum xy}{\sqrt{\sum x2 \sum y2}}$
- Correlation coefficient r lies between –1 and 1. (i.e) $-1 \leq r \leq 1$
- When $r=+1$, then the correlation is perfect positive
- When $r=-1$, then the correlation is perfect negative
- When $r=0$, then there is no relationship between the variables, (i.e) the variables are uncorrelated.
- Spearman's Rank correlation deals with qualitative characteristics.

9.11 KEY WORDS

Correlation , Spearman's Rank correlation, Pearson correlation, Correlation Coefficient , Scattered Diagram

9.12 ANSWER TO CHECK YOUR PROGRESS

1. The term correlation refers to the degree of relationship between two or more variables
2. Linear correlation is a measure of the degree to which two variables vary together, or ameasure of the intensity of the association between two variables
3. Positive and Negative correlation ,Simple, Partial and Multiple correlation,Linear and Non-Linear correlation
4. Scatter diagram is a graphic device for finding correlation between two variables
5. $r_s = 1 - \frac{6\sum D^2}{n(n^2-1)}$

9.13 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. Calculate the coefficient of correlation from the following data: Σ*X*=50, Σ*Y*=–30, Σ*X*2 =290, ΣY2 =300, Σ*XY*=–115, *N*=10
2. The following data pertains to the marks in subjects *A* and *B* in a certain examination. Mean marks in *A* = 39.5, Mean marks in *B*= 47.5 standard deviation of marks in *A* =10.8 and Standard deviation of marks in *B*= 16.8. coefficient of correlation between marks in *A* and marks in *B* is 0.42. Give the estimate of marks in B for candidate who secured 52 marks in A.
3. What is scattered diagram and explain it?

LONG ANSWER QUESTIONS

1. A random sample of recent repair jobs was selected and estimated cost, actual cost were recorded. Calculate the value of Spearman's correlation

Estimated cost	70	68	67	55	60	75	63	60	72
Actual cost	65	65	80	60	68	75	62	60	70

2. Distinguish between Karl Pearson's coefficient and Spearman's correlation coefficient
3. Explain the types of correlation with examples

9.14 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers andDistributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing CompanyLtd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw HillPublishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., NewDelhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons.,NewDelhi.

6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., Delhi.

UNIT X - SPEARMAN'S RANK CORRELATION

Structure

10.0 INTRODUCTION

Regression means stepping back or going back. It was first used by Francis Galton in 1877. He studied the relationship between the height of father and their sons. The study revealed that

- Tall fathers have tall sons and short fathers have short sons.
- The mean height of the sons of tall father is less than mean height of their fathers.
- The mean height of sons of short fathers is more than the mean height of their fathers.

The tendency to going back was called by Galton as 'Line of Regression'. This line describing the average relationship between two variables is known as the line of Regression.

In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modelling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine

learning.

10.1 OBJECTIVES

After studying this chapter students will be able to understand

- Concept of Regression and Regression coefficients
- Types of regression equations
- Regression lines both x on y and y on x

10.2 REGRESSION

Regression analysis refers to assessing the relationship between the outcome variable and one or more variables. The outcome variable is known as the dependent or response variable and the risk elements, and cofounders are known as predictors or independent variables. The dependent variable is shown by "y" and independent variables are shown by "x" in regression analysis.

10.3 LINEAR REGRESSION

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable. For example, a modeler might want to relate the weights of individuals to their heights using a linear regression model.

10.4 TYPES OF REGRESSION EQUATIONS

The Regression Equation is the algebraic expression of the regression lines. It is used to predict the values of the dependent variable from the given values of independent variables. As there are two regression lines, there are two regression equations. For the two variables X and Y, there are two regression equations. They are.

- **Regression equation of X on Y.**
- **Regression equation of Yon X.**

10.4.1 Regression Equation of X on Y

The straight line equation is $X=a+by$

Here *a* and*b* are unknown constants, which determines the position. The constant *a* is the intercept on the other value; the constant b is the slope; the following two normal equations are derived;

$$\sum x = na + b\sum y$$

$$\sum xy = a\sum x + b\sum y^2$$

The Regression equation X on Y is used to find out the values of X for given value of Y.

10.4.2 Regression Equation of Y on X

The straight line equation is **Y=a+bx**

The following two normal equationsare derived

$$\sum y = na + b\sum x$$

$$\sum xy = a\sum x + b\sum x^2$$

The Regression equation Y on X is used to ascertion the value of y for a given value of x.

Example:

Find out the regression equation, x on y and y on x from the following data:

X	15	20	25	30	35	40	45
y	8	14	20	26	32	38	44

Solutions

x	**y**	**X^2**	**Y^2**	**xy**
15	8	225	64	120
20	14	400	196	280
25	20	625	400	500
30	26	900	676	780
35	32	1225	1024	1120
40	38	1600	1444	1520
45	44	2025	1936	1960
$\sum x=210$	**$\sum y=182$**	**$\sum x^2=7000$**	**$\sum y^2=5740$**	**$\sum xy=6300$**

$\sum x=210$; $\sum y=182$; $\sum x^2=7000$; $\sum y^2=5740$; $\sum xy=6300$

Regression equation x on y is $y = a + by$

Hence $\sum x = na + b\sum y$

$\sum xy = a\sum x + b\sum y^2$

$210 = 7a + 182b$

(1)

6300=182a+5740b (2)

Multiplying equation (1) by 26

5460=182a+4732b

(3)

6300 = 182a + 5740b

(4)

Deducting (3) from Equation (2)

6300=182a+5740b

(4)

5460=182a+4732b

(3)

(-) (-) (-)

840=0 + 1008b

Therefore, b=840/1008=0.83

Substituting the value of b in Eq.(1)

210=7a+(182x0.83)

210=7a+151.06

7a + 151.06=210

7a=210-151.06

7a=58.94

a= 8.42

Hence, x=a+by

x=8.42 + 0.83 y

Regression Eq of y on x y=a+bx

Hence, $\sum y = Na + b\sum y^2$

182=7a+210b

(1)

6300=210a+700b

(2)

Multiplying Eq.(1) by 30

5460=210a+6300b

(3)

6300=210a+7000b

(4)

Deducting Eq.(4)from Eq.(3)

6300=210a+7000b

(3)

5460=210a+6300b

(4)

840=0 +700b

700=840

b=840/700

=1.2

Substituting the value of b in Eq.(1)

182=7a+(120x1.2)

182=7a+252

7a+252=182

7a=182-252

7a= -70

a= -10

Therefore, y= -10+1.

10.5 CURVE FITTING BY THE METHOD OF LEAST SQUARE

1. Curve Fitting

Curve fitting is the process of introducing mathematical relationships between dependent and independent variables in the form of an equation for a given set of data.

2. Method of Least Squares

The method of least squares helps us to find the values of unknowns a and b in such a way that the following two conditions are satisfied:

- The sum of the residual (deviations) of observed values of Y and corresponding expected (estimated) values of Y will be zero. $\sum(Y-\hat{Y})=0$ $\sum(Y-Y^{\wedge})=0$.
- The sum of the squares of the residual (deviations) of observed values of YY and corresponding expected values (YˆY^) should be at least $\sum(Y-\hat{Y})2\sum(Y-Y^{\wedge})^2$.

3. Fitting of a Straight Line:

A straight line can be fitted to the given data by the method of least squares. The equation of a straight line or least square line isY=a+bX, where a and b are constants or unknowns.

To compute the values of these constants we need as many equations as the number of constants in the equation. These equations are called normal equations. In a straight line there are two constants a and b so we require two normal equations.

Normal Equation for '*a*' **∑Y = na+ b∑X**

Normal Equation for '*b*' **∑XY = a∑X+ b∑X2**

Example:

The given example explains how to find the equation of a straight line or a least square line by using the method of least square, which is very useful in statistics as well as in mathematics.

X	1	2	3	4	5
y	2	5	3	8	7

Solution

X	Y	XY	X^2	1.1+1.3X	Y- $\hat{Y}$
1	2	2	1	2.4	-0.4
2	5	10	4	3.7	1.3
3	3	9	9	5.0	-2
4	8	32	16	6.3	1.7
5	7	35	25	7.6	-0.6
∑X=15	∑Y=25	∑XY=88	$\sum X^2$=55	Trend values	∑(Y- $\hat{Y}$ =)

The equation of least square line Y=a + bx

Normal equation for '*a*' ∑Y=na+ b ∑X 25

= 5a + 15b∑Y = na+b∑X 25=5a+15b ---- (1)

Normal equation for '*b* $\sum XY=a\sum X+b\sum X2$ $88=15a+55b$ ----*(2)*

Eliminate a from equation (1) and (2),multiply equation (2) by 3 and subtract from equation (2). Thus we get the values of a and b.

Here a=1.1 and b=1.3X

For the trends values, put the values of X in the above equation

10.6 DERIVATIONS OF REGRESSION EQUATION

1. When deviations are taken from Arithmetic means of X and Y

The above method of finding out the regression equation is difficult. Instead, we can use the deviations of X and Y observations from their respective averages.

(i) **Regression equation of X on Y**

The regression equation of Y on X can also be expressed in the following form-

$$\mathbf{X - \bar{X} = r\ \sigma x\ /\sigma y\ (Y-\bar{Y})}$$

Here, $\bar{X}$ is the average of X observations and $\bar{Y}$ is the average of Y observations.

$r\sigma x/\sigma y$ is the regression coefficient of X on Y and is denoted by b_{xy}. b_{xy} measures the amount of change in X corresponding to a unit change in Y.

$$\mathbf{r\ \sigma x/\sigma y = bxy = \sum xy/\sum y^2}$$

Where $x=(X-\bar{X})$ and $y=(Y-\bar{Y})$

(ii) **Regression equation of Y on X**

The regression equation of Y on X can also be expressed in the following form-

$$\mathbf{Y - \bar{Y} = r\ \sigma y/\sigma x(X - \bar{X})}$$

$r\sigma y/\sigma x$ is the regression coefficient of Y on X and is denoted by b_{yx}. byx measures the amount of change in Y corresponding to a unit change in X.

$$\mathbf{r\ \sigma y/\sigma x = b_{yx} = (\sum xy)/\sum x^2}$$

We can calculate the coefficient of correlation which is the geometric mean of the two regression coefficients (b_{xy} &b_{yx}) i.e.

$$\mathbf{r = \sqrt{(b_{xy})\times(b_{yx})}}$$

2. When deviations are Taken from Assumed Mean

When instead of using actual means of X and Y observations, we use any arbitrary item (in the observation) as the mean.

We consider taking deviations of X and Y values from their respective assumed means.

The formula for calculating regression coefficient when regression is Y on X is as follows:

$$r\frac{\sigma y}{\sigma x} = b_{yx} = \frac{\sum(dx)(dy) - \frac{(\sum dx)(\sum dy)}{N}}{\sum(dx)\wedge 2 - \frac{(\sum dx)\wedge 2}{N}}$$

Where dx = (X-A_x) {A_x = assumed mean of X observations} and dy = (Y-A_y) {A_y = assumed mean of Y observations}

The formula for calculating regression coefficient when regression is X on Y is as follows:

$$r\frac{\sigma x}{\sigma y} = b_{xy} = \frac{\sum(dx)(dy) - \frac{(\sum dx)(\sum dy)}{N}}{\sum(dy)\wedge 2 - \frac{(\sum dy)\wedge 2}{N}}$$

In the case of **Grouped frequency distribution**, the regression coefficients are calculated from the bivariate frequency table (or correlation table).

The formula for calculating regression coefficient (in case of grouped frequency distribution) when regression is of Y on X is as follows-

$$r\frac{\sigma y}{\sigma x} = b_{yx} = \frac{\sum f(dx)(dy) - \frac{(\sum fdx)(\sum fdy)}{N}}{\sum(fdx)\wedge 2 - \frac{(\sum fdx)\wedge 2}{N}} \times \frac{hy}{hx}$$

Where hx= class interval of X variable and hy = class interval of Y variable

The formula for calculating regression coefficient (in case of grouped frequency distribution) when regression is of X on Y is as follows-

$$r\frac{\sigma x}{\sigma y} = b_{xy} = \frac{\sum f(dx)(dy) - \frac{(\sum fdx)(\sum fdy)}{N}}{\sum(fdy)\wedge 2 - \frac{(\sum fdy)\wedge 2}{N}} \times \frac{hx}{hy}$$

Check your Progress - 1

1. What are regression coefficients?
2. What is the formula used to calculate assumed mean?
3. What is method of least square?

10.7 PROPERTIES OF REGRESSION EQUATION

The constant 'b' in the regression equation (Y_e = a + bX) is called as the **Regression Coefficient**. It determines the slope of the line, i.e. the change in the value of Y corresponding to the unit change in X and therefore, it is also called as a **"Slope Coefficient."**

1. The correlation coefficient is the geometric mean of two regression coefficients. Symbolically, it can be expressed as:

$$r = \sqrt{b_{xy} + b_{yx}}$$

2. The value of the coefficient of correlation **cannot exceed unity i.e. 1.** Therefore, if one of the regression coefficients is greater than unity, the other must be less than unity.
3. The sign of both the regression coefficients will be same, i.e. they will be either positive or negative. Thus, it is not possible that one regression coefficient is negative while the other is positive.
4. The coefficient of correlation will have the same sign as that of the regression coefficients, such as if the regression coefficients have a positive sign, then "r" will be positive and vice-versa.
5. The average value of the two regression coefficients will be greater than the value of the correlation. Symbolically, it can be represented as

$$\frac{b_{xy} + b_{yx}}{2} > r$$

6. The regression coefficients are independent of the change of origin, but not of the scale. By origin, we mean that there will be no effect on the regression coefficients if any constant is subtracted from the value of X and Y. By scale, we mean that if the value of X and Y is either multiplied or divided by some constant, then the regression coefficients will also change.

   Thus, all these properties should be kept in mind while solving for the regression coefficients.

10.8 SUMMARY

- Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data
- Regression analysis refers to assessing the relationship between the outcome variable and one or more variables
- A straight line can be fitted to the given data by the method of least squares
- The constant 'b' in the regression equation (Y_e = a + bX) is called as the **Regression Coefficient**. It determines the slope of the line, i.e. the change in the value of Y corresponding to the unit change in X and therefore, it is also called as a **"Slope Coefficient."**

10.9 KEY WORDS

Regression, Linear regression, Types of regression coefficient, Properties of regression coefficient , straight line, Regression equation, straight line Deviations, Actual Mean,

10.10 ANSWER TO CHECK YOUR PROGRESS

1. Regression analysis refers to assessing the relationship between the outcome variable and one or more variables

2. When regression is Y on X

$$r\frac{\sigma y}{\sigma x} = b_{yx} = \frac{\sum(dx)(dy) - \frac{(\sum dx)(\sum dy)}{N}}{\sum(dx)\wedge 2 - \frac{(\sum dx)\wedge 2}{N}}$$

When regression is X on Y

$$r\frac{\sigma x}{\sigma y} = b_{xy} = \frac{\sum(dx)(dy) - \frac{(\sum dx)(\sum dy)}{N}}{\sum(dy)\wedge 2 - \frac{(\sum dy)\wedge 2}{N}}$$

3. The method of least squares helps us to find the values of unknowns a and b in such a way that the following two conditions are satisfied:

    - **∑(Y−Yˆ)=0 ∑(Y−Y^)=0.**
    - **∑(Y−Yˆ)2∑(Y−Y^)^2.**

10.10 QUESTIONS AND EXERCISE

SHORT QUESTION ANSWER

1. What are regression coefficients?
2. Define regression and write down the two regression equations
3. Describe different types of regression
4. What are the uses of regression analysis**?**

LONG QUESTION AND ANSWER

1. Explain the principle of least squares
2. State the properties of regression equations
3. For 5 observations of pairs of (X, Y) of variables X and Y the following results are

obtained. $\Sigma X=15$, $\Sigma Y=25$, $\Sigma X2=55$, $\Sigma Y2=135$, $\Sigma XY=83$. Find the equation of the
lines of regression and estimate the values of X and Y if Y=8 ; X=12.

4. Using the following information you are requested to (i) obtain the linear regression
of Y on X (ii) Estimate the level of defective parts delivered when inspectionexpenditure amounts to Rs.28,000 $\Sigma X=424$, $\Sigma Y=363$, $\Sigma X2 =21926$, $\Sigma Y2 =15123$,$\Sigma XY=12815$, N=10. Here X is the expenditure on inspection, Y is the defectiveparts delivered

10.11 FURTHER READINGS

1. Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers andDistributors (P) Ltd., Agra.
2. Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing CompanyLtd., New Delhi.
3. Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw HillPublishing Company Ltd., New Delhi.
4. Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., NewDelhi.
5. Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons., NewDelhi.
5. 6. Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.

UNIT – IX BUSINESS FORECASTING

Structure

11.1 Introduction

Business forecasting is a method to predict the future, where the future is narrowly defined by economic conditions. It combines information gathered from past circumstances with an accurate picture of the present economy to predict future conditions for a business.

It refers to techniques such as taking a prospective view of how the economy is likely to turn out in the short-term. Its use is critical for businesses whenever the future is uncertain. The more they can focus on the probable outcome, the more success the organization has as it moves forward.

11.2 The Objectives of Forecasting

In the narrow sense, the objective of forecasting is to produce better forecasts. But in the broader sense, the objective is to improve organizational performance—more revenue, more profit, increased customer satisfaction. Better forecasts, by themselves, are of no inherent value if those forecasts are ignored by management or otherwise not used to improve organizational performance.

A wonderfully sinister way to improve forecast accuracy (while ignoring more important things like order fill, customer satisfaction, revenue generation, and profit) was provided by Ruud Teunter of Lancaster University, at the 2008 International Symposium on Forecasting. Teunter compared various forecasting methods for a data set of 5,000 items having intermittent demand patterns. (Intermittent patterns have zero demand in many or most time periods.)

Teunter found that if the goal is simply to minimize forecast error, then forecasting zero in every period was the best method to use! (The zero forecast had lower error than a moving average, exponential smoothing, bootstrapping, and three variations of Croston's method that were tested.) However, for proper inventory management to serve customer needs, forecasting zero demand every period is probably not the right thing to do.

A similar point was made last fall in a Foresight article by Stephan Kolassa and Roland Martin (discussed in "Tumbling Dice"). Using a simple dice tossing experiment, they showed the implications for bias in commonly used percentage error metrics. What makes this important to management is that if the sole incentive for forecasters is to minimize MAPE, the forecaster could do best by purposely forecasting too low. This, of course,

could have bad consequences for inventory management and customer service.

11.3 Prediction, projection and forecasting

Forecast is scientific and free from intuition and personal bias, whereas prediction is subjective and fatalistic in nature.
Forecasting is an extrapolation of past into the future while prediction is judgmental and takes into account changes taking place in the future.
Therefore, prediction is utilized more in business and economics while forecasting takes place in weather and earthquakes.
Predicting is saying or telling something before the event while forecasting is done on the basis of analysis of the past.
Forecasting is still not a complete science as there are chances of error.
Concept of Forecasting
Forecasting is a process of making predictions about the future course of a business or a company based on trend analysis and past and present data.

So essentially data is collected and studied about the business, and analysis is done to forecast future scenarios that are likely to occur. Hence forecasting is an important tool in the process of business planning.

Forecasts are usually done by managers (at different levels, Statisticians, experts, economists, consultants etc. They involve a lot of data collecting (both past and present) and data analysis.
There is also the use of scientific techniques and methods. But at the end of the day, forecasting is not an exact science. There is always some guessing and observations involved. This is when the experience and knowledge of these experts come into play.

11.4 characteristics of forecasting are as follows:

Forecasting is strictly concerned with future events only
It analysis the probability of a future event or transaction occurring or happening
It involves analysis of data from the past and the present
Forecasting uses scientific techniques and methods to make such forecasts
But it also involves certain guesswork and observations

11.5 Steps in Forecasting

Identifying and Understanding the Structure
There are almost indefinite factors that can affect the future of a business. Identifying all these factors is neither possible nor desirable. So to make an accurate forecast, the managers have to identify the factors on which to focus. So internal and external factors must be studied to identify the strategic factors of the business.

Forecasting the future
Now that the foundation is laid the next step is to make an accurate and scientific forecast. This involves both scientific tools and techniques and also professional judgment and observations. The forecast is not a foolproof plan, only a guiding map for the future.

Analysis of Deviations

No forecast will be completely accurate. The differences or deviations from the forecasts should be analyzed and studied. This will help in making more accurate forecasts in the future.

Adapting the Forecasts Procedure
In forecasting, the skill and professional judgement required are gained by experiences and practice. The forecast process is fine-tuned with every cycle. So we can learn from our mistakes and shortcomings and keep improving on the forecasting procedures.

11.6 Methods of Business Forecasting

business forecasting. The methods are: 1. Bottom-up Method 2. Top-down Method 3. Historical Method 4. Deductive Method 5. Joint Opinion Method 6. Scientific Business Forecasting.

Business Forecasting: Method # 1.
Bottom-up Method:

Under this method various departments of an enterprise collect their own information/data and prepare their own forecasts. On the basis of these forecasts, the forecast for the firm as a whole is then undertaken. Thus, the responsibility of successful forecasting lies directly with various departments and people in the organisation.

Business Forecasting: Method # 2.
Top-down Method:

This method is just reverse of the direct or bottom-up method. In this method the forecast for the industry/business as a whole is ascertained first and then the particular forecasts for the various activities of the business are established. The process of forecasting is, thus, indirect and the responsibility for success in forecasting mainly lies with the top levels of management.

Business Forecasting: Method # 3.
Historical Method:

This method refers to the projection of trends on the basis of past events. The historical sequence of events is analysed as a basis for understanding the present situation and forecasting the future trends. The past recurring trends are associated with the corresponding cause and effect phenomenon in the future.

The important advantages of this method include:

(i) The past information or records can be easily obtained; and

(ii) Present information is also not ignored.
However, the main limitation of this method is that the future trends may deviate drastically from the normal path indicated by the past events. Further, it may not be possible to find trend or develop correlation between

cyclic movements of past data and other variables which have bearing upon them.

Business Forecasting: Method # 4.

Deductive Method:

Under this method future trends are based on observation and investigation. In addition to the critical analysis of the past events to draw future inferences, the subjective evaluation and conclusions for deducing discretion, experience and intuition of the forecaster.

This method can be regarded as more dynamic in character as it takes into consideration not only the historical sequence of events but also the latest developments. However, the main drawback of this method is that it relies more on individual judgement and initiative appraisal than on actual record.

Business Forecasting: Method # 5.

Joint Opinion Method:

As the name suggests, this method utilises the collective opinion, judgement and experience of various experts. A committee for business forecasting is formulated to take the joint view of various members. An attempt is made to evolve consensus for predicting future events on the basis of their views.

The main advantages of this method include:

(i) It encourages co-operation and co-ordination and also utilises the services of various experts;

(ii) There is no need of detailed statistical analysis, and

(iii) It is simple and easy to operate.

However, the main disadvantage of this method is the joint responsibility which may ultimately result into no-body's responsibility. The members of the committee may also not take active interest as they know that their judgement may not be finally accepted. This may degenerate the entire forecasting process into a mere guess work.

Business Forecasting: Method # 6.

Scientific Business Forecasting:

Under this method, forecasting is done on scientific lines by making use of various statistical tools, such as, business index or barometer, extrapolation or mathematical projections, regression and econometric models. Past statistical data modified in the light of changed present conditions provides the basic raw material for drawing more accurate conclusions for the future.

The following are some of the most important statistical tools used for business forecasting:

(a) Business Index or Barometer.

(b) Extrapolation or Mathematical Projection.

(c) Regression

(d) Econometric Model.

(a) Business Index or Barometer:

The term ‘business index’ refers to a series relating to business conditions. It is also known as ‘barometer’, ‘indicator’ or ‘economic forecaster.’ Such a business index number may relate to general conditions of business or to a particular trade or industry or to an individual business.

The index number may measure changes in business activity during the changes of cyclical variations, i.e. boom, decline, depression and recovery. It is called business barometer because it helps in making forecasts for future business conditions.

The indices of production, wages, trade, finance, stocks and shares, etc. are plotted on a graph paper to obtain the curve showing trend of long-period and seasonal movements. The various index numbers relating to different activities of business may be combined into a general or composite index of business activity.’

This general index is an indicator of future conditions of trade and industry in general. However, the behaviour of individual trade or industry might show a different trend from that of general index, As such, the study of general index should be supplemented by separate studies of individual trade or industry.

The following are some of the important series which are considered by businessmen for forecasting:

(i) Index of Wholesale Prices

(ii) Index of Consumer Prices

(iii) Index of Industrial Product

(iv) Gross National Product

(v) Employment

(vi) General Aggregate Consumption

(vii) Volume of Agricultural Production

(viii) Stock Exchange Index.

The different figures may be converted into relatives on a certain base. The weighted average of these relatives may be computed to ascertain the business index called the barometer.

These business barometers guide the businessmen in taking decisions on many problems like expansion of production activity, diversification, undertaking of a new project, exploring new markets, launching as sales campaign rising of funds through issue of shares or debentures etc.

The reports on general business and trade conditions are published by the Chamber of Commerce, industry and some trade associations. Important journals and newspapers also publish index numbers relating to various industries and trades. The Reserve Bank of India also publishes various index numbers and indicators of general economic conditions.

The business barometers are very useful in business forecasting, but sometimes these barometers give misleading conclusions due to inaccurate construction of index numbers or changed conditions. As many factors may prevent history to repeat it, it is necessary to modify the trend revealed by business barometers in the light of specific conditions influencing the judgement.

(b) Extrapolation or Mathematical Projection:

Extrapolation is the process of estimating a value for some future period, based on some assumptions.

The basic assumptions underlying this statistical tool of business forecasting include:

(i) There should not be sudden jumps in figures from one period to another; and

(ii) The conditions in the future will not change materially.

(c) Regression:

The regression equation, $y=a+bx$, can be used as an instrument to predict the value of y for a given value of x. The regression equation is highly used in physical sciences where the data are related functionally. However, in business forecasting it may be very difficult to establish functional relationships and hence the use of regression equation is also limited.

(d) Econometric Models:

Economic activities describe in terms of mathematical equations are referred to as econometric models. These models show the way of inter-relationships amongst the various aspects of the economy. The econometric models are not very popular because it is not possible for every business to develop his own model of the economy.

UNIT – XII TIME SERIES ANALYSIS

Structure

12.1 Introduction

A series of observations, on a variable, recorded after successive intervals of time is called a time series. The successive intervals are usually equal time intervals, e.g., it can be 10 years, a year, a quarter, a month, a week, a day, and an hour, etc. The data on the population of India is a time series data where time interval between two successive figures is 10 years. Similarly figures of national income, agricultural and industrial production, etc., are available on yearly basis.

12.2 Regression analysis

The main objective of regression analysis is to know the nature of relationship between two variables and to use it for predicting the most likely value of the dependent variable corresponding to a given, known value of the independent variable. This can be done by substituting in *Eq.(5.1a)* any known value of X corresponding to which the most likely estimate of Y is to be found.

For example, the estimate of Y (*i.e.* Y_c), corresponding to $X = 15$ is

$Y_c = 8.61 + 0.71(15)$

8.61 + 10.65
19.26

It may be appreciated that an estimate of Y derived from a regression equation will not be exactly the same as the Y value which may actually be observed. The difference between estimated Y_c values and the corresponding observed Y values will depend on the extent of scatter of various points around the line of best fit.

loser the various paired sample points *(Y, X)* clustered around the line of best fit, the smaller the difference between the estimated Y_c and observed Y values, and vice-versa. On the whole, the lesser the scatter of the various points around, and the lesser the vertical distance by which these deviate from the line of best fit, the more likely it is that an estimated Y_c value is close to the corresponding observed Y value.

12.3 Exponential Smoothing Method

The exponential smoothing method also facilities continuous updating of the estimate of MAD. The current MAD_t is given by
MAD_t = α Actual values- Forecasted values + (1-α) MAD_{t-1}

Higher values of smoothing constant α make the current MAD more responsive to current forecast errors.

Example 7.7: A firm uses simple exponential smoothing with α =0.1 to forecast demand. The forecast for the week of February 1 was 500 units whereas actual demand turned out to be 450 units.

Forecast the demand for the week of February 8.
Assume the actual demand during the week of February 8 turned out to be 505 units. Forecast the demand for the week of February 15. Continue forecasting through March 15, assuming that subsequent demands were actually 516, 488, 467, 554 and 510 units. Solution:
Given F_{t-1} = 500, D_{t-1} = 450, and α = 0.1
$F_t = F_{t-1} – α(D_{t-1} - F_{t-1})$ = 500 + 0.1(450-500) = 495 units

Forecast of demand for the week of February 15 is shown in Table 7.5

Table 7.5: Forecast of Demand

Week	Demand	Old	Forecast	Correction	New Forecast (F_t)
	D_{t-1}	Forecast	Error	$α(D_{t-1} - F_{t-1})$	$F_{t-1} + α(D_{t-1} - F_{t-1})$
		F_{t-1}	$(D_{t-1} – F_{t-1})$		
Feb. 1	450	500	-50	-5	495
Feb. 8	505	495	10	1	496
Feb. 15	516	496	20	2	498
Feb. 22	488	498	-10	-1	497
Mar. 1	467	497	-30	-3	494
Mar. 8	554	494	60	6	500
Mar. 15	510	500	10	1	501

If no previous forecast value is known, the old forecast starting point may be estimated or taken to be an average of some preceding periods.

The estimated Y_C values will coincide the observed Y values only when all the points on the scatter diagram fall in a straight line. If this were to be so, the sales for a given marketing expenditure could have been estimated with l00 percent accuracy. But such a situation is too rare to obtain. Since some of the points must lie above and some below the straight line, perfect prediction is practically non-existent in the case of most business and economic situations.

This means that the estimated values of one variable based on the known values of the other variable are always bound to differ. The smaller the difference, the greater the precision of the estimate, and vice-versa. Accordingly, the preciseness of an estimate can be obtained only through a measure of the magnitude of error in the estimates, called the ***error of estimate***.

12.4 Theories of Business Forecasting:

- ✓ Theory of Economic Rhythm
- ✓ Action and Reaction Approach
- ✓ Sequence Method or Time Lag Method
- ✓ Specific Historical Analogy
- ✓ Cross-Cut Analysis
- ✓ Model Building Approach

Business Forecasting: Theory # 1.

12.5 Theory of Economic Rhythm:

This theory propounds that the economic phenomena behave in a rhythmic manner and cycles of nearly the same intensity and duration tend to recur. According to this theory, the available historical data have to be analysed into their components, i.e. trend, seasonal, cyclical and irregular variations.

The secular trend obtained from the historical data is projected a number of years into the future on a graph or with the help of mathematical trend equations. If the phenomena is cyclical in behaviour, the trend should be adjusted for cyclical movements.

When the forecast for a year is to be split into months or quarters then the forecaster should adjust the projected figures for seasonal variations also with the help of seasonal indices.

It becomes difficult to predict irregular variations and hence, rhythm method should be used along with other methods to avoid inaccuracy in forecasts. However, it must be remembered that business cycles may not be strictly periodic and the very assumptions of this theory may not be true as history may not repeat.

Business Forecasting: Theory # 2.

12.6 Action and Reaction Approach:

This theory is based on the Newton's 'Third Law of Motion', i.e., for every action there is an equal and opposite reaction. When we apply this law to business, it implies that it there if depression in a particular field of business, there is bound to be boom in it sooner or later. It reminds us of the business, cycle which has four phases, i.e., prosperity, decline, depression and prosperity.

This theory regards a certain level of business activity as normal and the forecaster has to estimate the normal level carefully. According to this theory, if the price of commodity goes beyond the normal level, it must come down also below the normal level because of the increased production and supply of that commodity.

Business Forecasting: Theory # 3.

12.7 Sequence Method or Time Lag Method:

This theory is based on the behaviour of different businesses which show similar movements occurring successively but not simultaneously. As such, this method takes into account time lag based on the theory of lead-lag relationship which holds good in most cases.

The series that usually change earlier serve as forecast for other related series. However, the accuracy of forecasts under this method depends upon the accuracy with which time lag is estimated.

Business Forecasting: Theory # 4.

12.8 Specific Historical Analogy:

This theory is based on the assumption that history repeats itself. It simply implies that whatever happened in the past under a set of circumstances is likely to happen in future under the same set of conditions.

Thus, a forecaster has to analyse the past data to select such period whose conditions are similar to the period of forecasting. Further, while predicting for the future, some adjustments may be made for the special circumstances which prevail at the time of making the forecasts.

Business Forecasting: Theory # 5.

12.9 Cross-Cut Analysis:

In this method of business forecasting, the combined effect of various factors is not studied, but the effect of each factor, that has a bearing on the forecast, is studied independently. This theory is similar to the Analysis of Time Series under the statistical methods.

Business Forecasting: Theory # 6.

12.10 Model Building Approach:

This approach makes use of mathematical equations for drawing economic models. These models depict the inter-relationships amongst the various factors affecting the economy or business. The expected values for dependent variables are then ascertained by putting the values of known variables in the model. This approach is highly mechanical and this can be rarely employed in business conditions.

12.11 Utility of Business Forecasting

Meaning and Definition:Business forecasting is an act of predicting the future economic conditions on the basis of past and present information. It refers to the technique of taking a prospective view of things likely to shape the turn of things in foreseeable future. As future is always uncertain,

there is a need of organised system of forecasting in a business.Thus, scientific business forecasting involves:
(i) Analysis of the past economic conditions and

(ii) Analysis of the present economic conditions; so as to predict the future course of events accurately.

In this regard, business forecasting refers to the analysis of the past and present economic conditions with the object of drawing inferences about the future business conditions. In the words of Allen, “Forecasting is a systematic attempt to probe the future by inference from known facts. The purpose is to provide management with information on which it can base planning decisions.

Leo Barnes observes, “Business Forecasting is the calculation of reasonable probabilities about the future, based on the analysis of all the latest relevant information by tested and logically sound statistical econometric techniques, as interpreted, modified and applied in terms of an executive’s personal judgment and social knowledge of his own business and his own industry or trade”.
In the words of C.E. Sulton, “Business Forecasting is the calculation of probable events, to provide against the future. It therefore, involves a ‘look ahead’ in business and an idea of predetermination of events and their financial implications as in the case of budgeting.”

According to John G. Glover, “Business Forecasting is the research procedure to discover those economic, social and financial influences governing business activity, so as to predict or estimate current and future trends or forces which may have a bearing on company policies or future financial, production and marketing operations.”

The essence of all the above definitions is that business forecasting is a technique to analyse the economic, social and financial forces affecting the business with an object of predicting future events on the basis of past and present information.
Steps of Forecasting:
The process of forecasting consists of the following steps, also described as elements of forecasting:
1. Developing the Basis:
The first step involved in forecasting is developing the basis of systematic investigation of economic situation, position of industry and products. The future estimates of sales and general business operations have to be based on the results of such investigation. The general economic forecast marks as the primary step in the forecasting process.
2. Estimating Future Business Operations:
The second step involves the estimation of conditions and course of future events within the industry. On the basis of information/data collected through investigation, future business operations are estimated. The

quantitative estimates for future scale of operations are made on the basis of certain assumptions.

3. Regulating Forecasts:

The forecasts are compared with actual results so as to determine any deviations. The reasons for his variations are ascertained so that corrective action is taken in future.

4. Reviewing the Forecasting Process:

Once the deviations in forecasts and actual performance are found then improvements can be made in the process of forecasting. The refining of forecasting process will improve forecasts in future.

Sources of Data Used In Business Forecasting:

Collection of data is a first step in any statistical investigation. It is the basis for any analysis and interpretations. Before collection of data, many questions shall occupy the mind of the manager. The manager must be able to answer these questions before task of collection is started.

12.12 Limitations of Business Forecasting:

Inspite of many advantages, some people regard business forecasting "as an unnecessary mental gymnastics and reject it as a sheer waste of time, money and energy."

The reason for the same lies in the fact that despite all precautions, an element of error is bound to creep in the forecasts and we cannot eliminate guesswork in forecasts. It is also felt that forecasting is influenced by the pessimistic or optimistic attitude of the forecaster.

It may not be possible to make forecasts with a pin-point accuracy. But, it still cannot undermine the importance of business forecasting. The management should first make use of statistical and econometric models in making forecasts and then apply collective experience, skill and objective judgement in evaluating the forecasts.

12.13 Business Forecasting: Advantage

Business Forecasting: Advantage # 1. Establishing a New Business:

While setting up a new business, a number of business forecasts are required. One has to forecast the demand for the product, capacity of competitors, expected share in the market, the amount and sources of raising finances, etc.

The success of a new business will depend upon the accuracy of such forecasts. If the forecasts are made systematically, then the operations of the business will go smoothly and the chances of failure will be minimised.

Business Forecasting: Advantage # 2. Formulating Plans:

Forecasting provides a logical basis for preparing plans. It plays a major role in managerial planning and supplies the necessary information. The future assessment of various factors is essential for preparing plans. In fact, planning without forecasting is an impossibility. Henry Fayol has rightly observed that the entire plan of an enterprise is made up of a series of plans called forecasts.

Business Forecasting: Advantage # 3. Estimating Financial Needs:

Every business needs adequate capital. In the absence of correct estimates of financial requirements, the business may suffer either from inadequate or from excess capital. Forecasting of sales and expenses helps in estimating future financial needs.

The plans for expansion, diversification or improvement also necessitate the forecasting of requirements of funds. A proper financial planning depends upon systematic forecasting.

Business Forecasting: Advantage # 4. Facilitating Managerial Decisions:
Forecasting helps management to take correct decisions. By providing a logical basis for planning and determining in advance the nature of future business operations, it facilitates correct managerial decisions about material, personnel, sales and other requirements.

Business Forecasting: Advantage # 5. Quality of Management:
It improves the quality of managerial personnel by compelling them to look into the future and make provision for the same. By focussing attention on future, forecasting helps the management in adopting a definite course of action and a set purpose.

Business Forecasting: Advantage # 6. Encourages Co-operation and co-ordination:
Forecasting calls for some minimum effort on the part of all and. thus, creates a sense of participation. It is not a one man's or one department's job. No department or person can make its forecasts in isolation. There should be a proper co-operation and co-ordination among different departments for setting proper forecasts for the business as a whole.

So, forecasting process leads to better co-operation and co-ordination among people of various departments of the organisation.

Business Forecasting: Advantage # 7. Better Utilisation of Resources:
Forecasting ensures better utilisation of resources by revealing the areas of weaknesses and providing necessary information about the future. Management can concentrate on critical areas and control more effectively.

Business Forecasting: Advantage # 8. Success in Business:
Success in business, to a great extent, depends upon correct predictions about the future. Systematic forecasting ensures smooth and continuous working of the business. By knowing the future course of events in advance, one could always face the difficulties in a planned manner.

UNIT XIII - ANALYSIS OF TIME SERIES

Structure

13.0 INTRODUCTION

When quantitative data are arranged in the order of their occurrence, the resulting statistical series is called a time series. The quantitative values are usually recorded over equal time interval daily, weekly, monthly, quarterly, half yearly, yearly, or any other time measure. Monthly statistics of Industrial Production in India, Annual birth-rate figures for the entire world, yield on ordinary shares, weekly wholesale price of rice, and daily records of tea sales or census data are some of the examples of time series. Each has a common characteristic of recording magnitudes that vary with passage of time. In this unit we will see about time series analysis.

13.1 OBJECTIVES

After going through this unit, you will

- Learn about time series analysis
- Know about the measurement of trends

- Understand forecasting and Deseasonalisation

13.2 TIME SERIES ANALYSIS

Time series are influenced by a variety of forces. Some are continuously effective other make themselves felt at recurring time intervals, and still others are non-recurring or random in nature. Therefore, the first task is to break down the data and study each of these influences in isolation. This is known as decomposition of the time series. It enables us to understand fully the nature of the forces at work. We can then analysis their combined interactions. Such a study is known as time-series analysis.

13.2.1 COMPONENTS OF TIME SERIES:

The factors that are responsible for bringing about changes in a time series, also called the components of time series, are as follows:

- Secular Trends (or General Trends)
- Seasonal Movements
- Cyclical Movements
- Irregular Fluctuations

Secular Trends:

Secular trend is the main component of a time series which results from long term effects of socio-economic and political factors. It shows the growth or decline in a time series over a long period. It is the type of tendency which continues to persist for a very long period. Prices and export and import data, for example, reflect obviously increasing tendencies over time.

Seasonal Trends:

Seasonal trends are short term movements occurring in data due to seasonal factors. The short term is generally considered as a period in which changes occur in a time series with variations in weather or festivities. For example, it is commonly observed that the consumption of ice-cream during summer is generally high and hence an ice-cream dealer's sales would be higher in some months of the year while relatively lower during winter months. Employment, output, exports, etc., are subject to change due to variations in weather. Similarly, the sale of garments, umbrellas, greeting cards and fire-works are subject to large variations during festivals like Valentine’s Day, Eid, Christmas, New Year's, etc. These types of variations in a time series are isolated only when the series is provided biannually, quarterly or monthly.

Cyclic Movements

It is a long term oscillations occurring in a time series. These

oscillations are mostly observed in economics data and the periods of such oscillations are generally extended from five to twelve years or more. These oscillations are associated with the well known business cycles. These cyclic movements can be studied provided a long series of measurements, free from irregular fluctuations, is available.

Irregular Fluctuations

It happens when a sudden changes occurring in a time series which are unlikely to be repeated. They are components of a time series which cannot be explained by trends, seasonal or cyclic movements. These variations are sometimes called residual or random components. These variations, though accidental in nature, can cause a continual change in the trends, seasonal and cyclical oscillations during the forthcoming period. Floods, fires, earthquakes, revolutions, epidemic, strikes etc., are the root causes of such irregularities.

13.2.2 ANALYSIS OF TIME SERIES

The objective of the time series analysis is to identify the magnitude and direction of trends, to estimate the effect of seasonal and cyclical variations and to estimate the size of the residual component. This implies the decomposition of a time series into its several components. Two lines of approach are usually adopted in analyzing a given time series:

- The additive model
- The multiplicative model

The additive model:

It is used when the four components of a time series are independent of one another. Independent means the magnitude and patterns of movement of the components do not affect each other. Using this assumption the magnitudes of the time series are regarded as the sum of separate influences of its four components. In additive approach, the unit of measurements remains the same for all the four components. The additive model can be written as

$$\mathbf{Y = T + S + C + R}$$

Where Y = magnitude of a time series

T = Trend,

C =Cyclical component,

S =Seasonal component,

R = Random component.

The multiplicative model:

It is used where the forces giving rise to the four types of

variations are interdependent. The magnitude of time series is the product of four components. Then the multiplicative model can be written as

$$\mathbf{Y = T \times S \times C \times R}$$

The additive model is usually used when the time series is spread over a short time span or where the rate of growth or decline in the trend is small. The multiplicative model, which is used more often than the additive model, is generally used whenever the time span of the series is large or the rate of growth or decline is

$$\mathbf{Y - T = S + C + R} \quad \textbf{or} \quad \frac{\mathbf{Y}}{\mathbf{T}} \mathbf{= S \times C \times R}$$

Similarly, a de-trended, de-seasonalized series may be obtained as

$$\mathbf{Y - T - S = C + R} \qquad \textbf{or} \qquad \frac{Y}{TxS} \mathbf{= C \times R}$$

It is not always necessary for the time series to include all four types of variations; rather, one or more of these components might be missing altogether. For example, when using annual data the seasonal component may be ignored, while in a time series of a short span having monthly or quarterly observations, the cyclical component may be ignored

13.3 MEASUREMENT OF TRENDS

- Moving average method
- Least square method

13.3.1 MOVING AVERAGE METHOD

Moving average method is a simple device of reducing fluctuations and obtaining rend values with a fair degree of accuracy. In this method the average value of a number of years (months, weeks, or days) is taken as the trend value for the middle point of the period of moving average. The process of averaging smoothes the curve and reduces the fluctuations.

The first thing to be decided in this method is the period of the moving average. What it means is to take a decision about the number of consecutive items whose average would be calculated each time. Suppose it has been decided that the period of the moving average would be 5 years (months, weeks, or days) then the arithmetic average of the first 2 items (number 1,2,34 and 5) would be placed against item no:3 and then the arithmetic average of item Nos:2,3,4,5 and 6would be placed against item No: 4. This process would be repeated till the arithmetic average of the last five items has been calculated.

Odd Period of Moving Average

Calculation of three yearly moving averages includes the following steps

1. Add up the values of the first 3 years and place the yearly sum against the median year. (This sum is called moving total)

2. Leave the first year value, add up the values of the next three years and place it against its median year.
3. This process must be continued till all the values of the data are taken for calculation.
4. Each 3-yearly moving total must be divided by 3 to get the 3-year moving averages, which is our required trend value.

   The formula calculating 3 yearly moving averages is as follows

   $\frac{a+b+c}{3}, \frac{b+c+d}{3}, \frac{c+d+e}{3}$

   The formula calculating 5 yearly moving averages is as follows

   $\frac{a+b+c+d+e}{5}, \frac{b+c+d+e+f}{5}, \frac{c+d+e+f+g}{5}$ …….

Example:

Calculate the 3 yearly and 5 yearly moving averages of the data

Years	1	2	3	4	5	6	7	8	9	10	11	12
sales	5.2	4.9	5.5	4.9	5.2	5.7	5.4	5.8	5.9	6.00	5.2	4.8

Solution:

Year	Sales	3 Year Moving Total	3 Year Moving Average (3) / 3	5 Year Moving Total	5 Year Moving Average (4) / 5
1	5.2	---		--	--
2	4.9	15.6	5.2	--	--
3	5.5	15.3	5.1	25.7	5.14
4	4.9	15.6	5.2	26.2	5.24
5	5.2	15.8	5.27	26.7	5.34
6	5.7	16.3	5.41	27.0	5.4
7	5.4	16.9	5.63	28.0	5.6
8	5.8	17.1	5.7	28.8	5.76
9	5.9	17.7	5.23	28.3	5.66
10	6.0	17.1	5.7	27.7	5.54
11	5.2	16.0	5.33	---	---
12	4.8	---	---	---	---

Even Period of Moving Average:

The period of moving average is 4,6, or 8, it is even number. The four yearly total cannot be placed against any year as median 2.5 is

between the second and the third year. So the total should be placed in between the 2nd and 3rd years. We must centre the moving average in order to place the moving average against the year

Steps to find even period of moving average:

1. Add up the values of the first 4 years and place the sum against the middle of 2nd and 3rd year. (This sum is called 4 year moving total)
2. Leave the first year value and add next 4 values from the 2nd year onward and write the sum against its middle position.
3. This process must be continued till the value of the last item is taken into account.
4. Add the first two 4-years moving total and write the sum against 3rd year.
5. Leave the first 4-year moving total and add the next two 4-year moving total and place it against 4th year.
6. This process must be continued till all the 4-yearly moving totals are summed up and centered.
7. Divide the 4-years moving total by 8 to get the moving averages which are our required trend values

Example:

Find the 4 yearly moving average foe determining trend values in the following time series data

Year	2005	2006	2007	2008	2009	2010	2011
Profit in(000) □	12	14	16	15	13	14	18

Solution:

Years	Profit	Sum of Fours	4 years Moving Average	4 yearly Moving Average Centered
2005	12			
2006	14			
		57	14.25	(14.25 + 14.50)/ 2 = 14.38
2007	16			
		58	14.50	(14.50 + 14.50)/ 2 = 14.50
2008	15			
		58	14.50	(14.50 + 15.00)/ 2 =

				14.75
2009	13			
		60	15.00	
2010	14			
2011	18			

Advantages

Moving averages can be used for measuring the trend of any series. This method is applicable to linear as well as non-linear trends.

Disadvantages

The trend obtained by moving averages generally is neither a straight line nor a standard curve. For this reason the trend cannot be extended for forecasting future values. Trend values are not available for some periods at the start and some values at the end of the time series. This method is not applicable to short time series

13.3.2 LEAST SQUARES METHOD

When the trend is linear the trend equation may be represented by y = a + bt and the values of a and b for the line y = a + bt which minimizes the sum of squares of the vertical deviations of the actual (observed) values from the straight line, are the solutions to the so called normal equations:

$$\Sigma y = na + b\Sigma t \quad \ldots\ldots\ldots\ldots\ldots.. (1)$$

$$\Sigma yt = a\Sigma t + b\Sigma t^2 \quad \ldots\ldots\ldots\ldots.(2)$$

Where n is the number of paired observations

The normal equation are obtained by multiplying y = a + bt, by the coefficient of a and b, i.e., by 1 and t throughout and summing up.

When the Number of Years is Odd

We can use this method when we are given odd number of years. It is easy and is widely used in practice. If the number of items is odd, we can follow the following steps:

1. Denote time as the t variable and values as y
2. Middle year is assumed as the period of origin and find out deviations
3. Square the time deviations and find t^2.
4. Multiply the given value of y by the respective deviation of t and find the total Σty.
5. Find out the values of y; get Σy
6. The value so obtained are placed in the two quations
   i. $\Sigma y = na + b\Sigma t$
   ii. $\Sigma yt = a\Sigma t + b\Sigma t^2$; find out the value of a and b
7. The calculated values of a and b are substituted and the trend

value of y are found for various values of t.

When the number of years is odd the calculation will be simplified by taking the mid year as origin and one year as unit and in that case

Σt = 0 and the two normal equations take the form

$\mathbf{\Sigma y = na \; ; \; \Sigma yt = b\Sigma t^2}$

Hence $a = \frac{\Sigma y}{n}$, $b = \frac{\Sigma yt}{\Sigma t^2}$

Example :

Calculate trend values by the method of least square from data given below and estimate the sales for 2003

Years:	1996	1997	1998	1999	2000
Sales of Co.A, (□ Lakhs)	70	74	80	86	90

Solution:

Year	**Sales**	**Deviation from 1998**		
	y	t	ty	t^2
1996	70	-2	-140	4
1997	74	-1	-74	1
1998	80	0	0	0
1999	86	1	86	1
2000	90	2	180	4
n = 5	**Σy = 400**	**Σt = 0**	**Σty = 52**	$\mathbf{\Sigma t^2 = 10}$

Since Σt = 0

$a = \frac{\Sigma y}{n} = \frac{400}{5} = 80$, $b = \frac{\Sigma yt}{\Sigma t^2} = \frac{52}{10} = 5.2$

Hence, y = 80 + 5.2 x t

Therefore $y_{1996} = 80 + 5.2\,(-2) = 80 - 10.4 = 69.6$

y_{1997} = 80 + 5.2 (-1) = 80 – 5.2 = 74.8

y_{1998} = 80 + 5.2 (0) = 80 + 0 = 80

y_{1996} = 80 + 5.2 (1) = 80 + 5.2 = 85.2

y_{1996} = 80 + 5.2 (2) = 80 + 10.4 = 90.4

For 2003, t will be 5. Putting t = 5 in the equation

Y_{2013} = 80 + 5.2 (5\0 = 80 + 26 = 106

Thus the estimated sales for the year 2003 is □ 106 lakhs

When the Number of Years is Even

When the number of years is even the origin is placed in the midway between the two middle years and the unit is taken to be half year instead of one year. With this change of origin and scale we have

$$\Sigma t = 0$$

$$\text{Hence } a = \frac{\Sigma y}{n}, \quad b = \frac{\Sigma yt}{\Sigma t^2}$$

Example:

Production of a company for 6 consecutive years is given in the following table. Calculate the trend value by using the method of least square

Year	2000	2001	2002	2003	2004	2005
Production	12	13	18	20	24	28

Solution:

Year	**Sales**	**Deviation from 2002.5**			Trend values
	y	t	ty	t^2	
2000	12	-2.5	-30	6.25	11.5
2001	13	-1.5	-19.5	2.25	14.5
2002	18	-0.5	-9	0.25	17.53
2003	20	0.5	10	0.25	20.81
2004	24	1.5	36	2.25	24.09

2005	28	2.5	70	6.25	27.37
n = 6	**Σy = 115**	**Σt = 0**	**Σty = 57.5**	**Σt^2 = 17.5**	

Since t = 0

$a = \frac{\Sigma y}{n} = \frac{115}{6} = 19.17$, $b = \frac{\Sigma yt}{\Sigma t^2} = \frac{57.5}{17.5} = 3.28$

Hence, y = 19.17 + 3.28 x t

Therefore y_{2000} = 19.17 + 3.28 (- 2.5) = 19.17 – 8.2 = 11.5

y_{2001} = 19.17 + 3.28 (- 1.5) = 19.17 – 4.92 = 14.5

y_{2002} = 19.17 + 3.28 (- 0.5) = 19.17 – 1.64 = 17.53

y_{2003} = 19.17 + 3.28 (0.5) = 19.17 + 1.64 = 20.81

y_{2004} = 19.17 + 3.28 (1.5) = 19.17 + 4.92 = 24.09

y_{2005} = 19.17 + 3.28 (2.5) = 19.17 + 8.2 = 27.37

Merits

1. The method is mathematically sound.
2. The estimates a and b are unbiased.
3. The least square method gives trend values for all the years and the method is devoid of all kinds of subjectivity.
4. The algebraic sum of deviations of actual values from trend values is zero and the sum of the deviations is minimum.

Demerits

1. The least square method is highly mathematical; therefore, it is difficult for a layman to understand it.
2. The method is not flexible.
3. It has been assumed that y is only a linear function of time period n. This may not be true in any situations.

13.4 MEASUREMENT OF SEASONAL VARIATION

Seasonal variations are that rhythmic changes in the time series data that is regular and periodic variations having a period of one year duration. Some of the examples which show seasonal variations are production of cold drinks, which are high during summer months and low during winter season. Sales of sarees in a cloth store which are high during festival season and low during other periods. They have their origin in climatic or institutional factors that affect either supply or demand or both. It is important that these variations should be measured accurately. The reason for determining seasonal variations in a time series is to isolate it and to study its effect on the size of the variable in the index form which is usually referred as seasonal index.

13.4.1 METHODS OF CONSTRUCTING SEASONAL INDICES

There are four methods of constructing seasonal indices.

1. Simple averages method
2. Ratio to trend method
3. Percentage moving average method
4. Link relatives method

Simple Average Method :

The time series data for each of the 4 seasons (for quarterly data) of a particular year are expressed as percentages to the seasonal average for that year. The percentages for different seasons are averaged over the years by using simple average. The resulting percentages for each of the 4 seasons then constitute the required seasonal indices.

Steps to calculate Simple Average Method:

(i) Arrange the data by months, quarters or years according to the data given.

(ii) Find the sum of the each months, quarters or year.

(iii) Find the average of each months, quarters or year.

(iv) Find the average of averages, and it is called Grand Average (G)

(v) Compute Seasonal Index for every season (i.e) months, quarters or year is given by

Seasonal Index (S.I) = $\frac{SeasonalAverage}{Grandaverage} \times 100$

If the data is given in months

Seasonal Index for Jan (S.I) = $\frac{\text{monthly } Average\ (forjan)}{Grandaverage} \times 100$

Seasonal Index for Feb (S.I) = $\frac{\text{monthly } Average\ (forfeb)}{Grandaverage} \times 100$

Similarly we can calculate SI for all other months

Example:

Calculate the seasonal index for the quarterly production of a computer using method of simple average

Year	I Quarter	II Quarter	III Quarter	IV Quarter
2011	355	451	525	500

2012	369	410	496	510
2013	391	432	458	495
2014	298	389	410	457
2015	300	390	431	459
2016	350	400	450	500

Solution:

Year	I Quarter	II Quarter	III Quarter	IV Quarter
2011	355	451	525	500
2012	369	410	496	510
2013	391	432	458	495
2014	298	389	410	457
2015	300	390	431	459
2016	350	400	450	500
Quarterly Total	2063	2472	2770	2921
Quarterly Averages	343.83	412	461.67	486.83

Seasonal Index (S.I) = $\frac{SeasonalAverage}{GrandAverage} \times 100$

Grand average = $\frac{343.83 + 412 + 461.67 + 486.83}{4} = \frac{1704.33}{4}$ = 426.0825

S.I for I Q = $\frac{343.83}{426.0825} \times 100$ = 80.69

S.I for II Q = $\frac{412}{426.0825} \times 100$ = 96.69

S.I for III Q = $\frac{461.67}{426.0825} \times 100$ = 108.35

S.I for IV Q = $\frac{486.83}{426.0825} \times 100$ = 114.26

Advantage and Disadvantage:

- Method of simple average is easy and simple to execute.
- This method is based on the basic assumption that the data do not contain any trend and cyclic components. Since most of the economic and business time series have trends and as such this method though simple is not of much practical utility.

CHECK YOUR PROGRESS - 1

1. A time series is a set of data recorded______________
2. The terms prosperity, recession, depression and recovery are in particular attached to ________
3. What is time series?

13.5 FORECASTING

Time series forecasting methods produce forecasts based solely on historical values and they are widely used in business situations where forecasts of a year or less are required. These methods used are particularly suited to Sales, Marketing, Finance, Production planning etc. and they have the advantage of relative simplicity. Time series forecasting is a technique for the prediction of events through a sequence of time.

The technique is used across many fields of study, from geology to economics. The techniques predict future events by analyzing the trends of the past on the assumption that future trends will hold similar to historical trends. Data is organized around relatively deterministic timestamps, and therefore, compared to random samples, may contain additional information that is tried to extract.

- Time series methods are better suited for short-term forecasts (i.e., less than a year).
- Time series forecasting relies on sufficient past data being available and that the data is of a high quality and truly representative.
- Time series methods are best suited to relatively stable situations. Where substantial fluctuations are common and underlying conditions are subject to extreme change, then time series

methods may give relatively poor results.

Advantages of forecasting:

1. Helps to predict the future:
2. Learns from the past
3. Remain competitive
4. Prepare for new business

Disadvantages of forecasting:

1. Basis of forecasting
2. Reliability of past data
3. Time and cost factor

13.6 DESEASONALISATION

When the seasonal component is removed from the original data, the reduced data are free from seasonal variations and is called deseasonalised data. That is, under a multiplicative model

$$\frac{\mathbf{TxSxCxI}}{\mathbf{S}} = \mathbf{T \ x \ C \ x \ I}$$

Deseasonalised data being free from the seasonal impact manifest only average valueof data.

Seasonal adjustment can be made by dividing the original data by the seasonal index.

$$\textbf{Deseasonalised data} = \frac{\boldsymbol{ORIGINALDATA}}{\boldsymbol{SEASONALINDEX}} \boldsymbol{X} \ \mathbf{100}$$

where an adjustment-multiplier 100 is necessary because the seasonal indices are usually given in percentages.

In case of additive model

$$\mathbf{Y_t = T + S + C + I}$$

$$\textbf{Deseasonalised data} = \boldsymbol{originaldata} - \frac{\boldsymbol{seasonalindex}}{\mathbf{100}}$$

$$= \mathbf{Yt} \ - \frac{\boldsymbol{seasonalindex}}{\mathbf{100}}$$

CHECK YOUR PROGRESS - 2

4. Define forecasting?
5. What are the method used for finding seasonal indices

13.7 SUMMARY

- Time series are influenced by a variety of forces. Some are continuously effective other make themselves felt at recurring

time intervals, and still others are non-recurring or random in nature. Therefore, the first task is to break down the data and study each of these influences in isolation. This is known as decomposition of the time series.

- The objective of the time series analysis is to identify the magnitude and direction of trends, to estimate the effect of seasonal and cyclical variations and to estimate the size of the residual component. This implies the decomposition of a time series into its several components. Two lines of approach are usually adopted in analyzing a given time series:
    - The additive model, the multiplicative model
- Moving average method is a simple device of reducing fluctuations and obtaining rend values with a fair degree of accuracy. In this method the average value of a number of years (months, weeks, or days) is taken as the trend value for the middle point of the period of moving average. The process of averaging smoothes the curve and reduces the fluctuations.
- When the trend is linear the trend equation may be represented by y = a + bt and the values of a and b for the line y = a + bt which minimizes the sum of squares of the vertical deviations of the actual (observed) values from the straight line, are the solutions to the so called normal equations:
- Seasonal variations are that rhythmic changes in the time series data that is regular and periodic variations having a period of one year duration.
- There are four methods of constructing seasonal indices. They are Simple averages method, Ratio to trend method, Percentage moving average method, Link relatives method.
- Time series forecasting methods produce forecasts based solely on historical values and they are widely used in business situations where forecasts of a year or less are required.
- When the seasonal component is removed from the original data, the reduced data are free from seasonal variations and is called deseasonalised data.

13.8 KEY WORDS

Time series,decomposition of the time series, additive model, the multiplicative model, Moving average method , least square method , Seasonal variations , Simple averages method, Ratio to trend method, Percentage moving average method, Link relatives method, forecasting , deseasonalised.

13.9 ANSWERS TO CHECK YOUR PROGRESS

1. Periodically, at equal time intervals, at successive points of time

2.Cyclical movements

3. Time series are influenced by a variety of forces. Some are continuously effective other make themselves felt at recurring time intervals, and still others are non-recurring or random in nature.

4. Time series forecasting methods produce forecasts based solely on historical values and they are widely used in business situations where forecasts of a year or less are required

5.There are four methods of constructing seasonal indices.

1. Simple averages method
2. Ratio to trend method
3. Percentage moving average method
4. Link relatives method

13.10 QUESTION AND EXERCISE

SHORT ANSWER QUESTION

1. What is time series?
2. What are the uses of time series
3. What are basic types of variations

LONG ANSWER QUESTION

1. Explain the components of time series
2. What are the various methods of estimating the trend components
3. Explain the moving average method? How is it calculated?
4. Describe the method of finding seasonal indices
5. Calculate the trend value by using three yearly moving average of the following data

Year	**1990**	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**	**1997**	**1998**	**1999**
Production	21	22	23	25	24	22	25	26	27	26

13.11 FURTHER READINGS

1. Spiegel, Murray R.: Theory and Practical of Statistics., London McGraw Hill Book Company.
2. Yamane, T.: Statistics: An Introductory Analysis, New York, HarperedRow Publication

3. R.P. Hooda: Statistic for Business and Economic, McMillan India Ltd.
4. G.C. Beri: Statistics for Mgt., TMH.
5. J.K. Sharma: Business Statistics, Pearson Education.
6. 6. S.P. Gupta : Statistical Methods, Sultan Chand and Sons.

UNIT XIV - INDEX NUMBER

Structure

14.0 INTRODUCTION

Index numbers are a commonly used statistical device for measuring the combined fluctuations in group-related variables. If we wish to compare the prices of consumer items today with their prices ten years ago, we are not interested in comparing the prices of only one item, but in comparing average price levels. We may wish to compare the present agricultural production or industrial production with that at the time of independence. Here again, we have to consider all items of production and each item may have undergone a different fractional increase (or even a decrease). How do we obtain a composite measure? This composite measure is provided by index numbers, which may be defined as a device for combining the variations that have occurred to a

group of related variables over a period of time, to obtain a figure that represents the 'net' result of the change in the constitute variables. In this unit you will learn in detail about index numbers.

14.1 OBJECTIVES

After going through this unit, you will

- Understand about index numbers and their types
- Learn about the different methods of calculating index numbers
- Know the uses and limitations of index numbers

14.2 INDEX NUMBERS

Index numbers are meant to study changes in the effects of factors which cannot be measured directly. According to Bowley, "Index numbers are used to measure the changes in some quantity which we cannot observe directly". For example, changes in business activity in a country are not capable of direct measurement, but it is possible to study relative changes in business activity by studying the variations in the values of some such factors which affect business activity, and which are capable of direct measurement.

Index numbers may be classified in terms of the variables that they are intended to measure. In business, different groups of variables in the measurement of which index number techniques are commonly used are (i) price, (ii) quantity, (iii) value and (iv)business activity. Thus, we have an index of wholesale prices, index of consumer prices, index of industrial output, index of value of exports and index of business activity, etc. Here we shall be mainly interested in index numbers of prices showing changes with respect to time, although the methods described can be applied to other cases. In general, the present level of prices is compared with the level of prices in the past. The present period is called the current period and some period in the past is called the base period.

14.2.1 TYPE OF INDEX NUMBER

Index numbers are names after the activity they measure. Their types are as under:

Price Index: Measure changes in price over a specified period of time. It is basically the ratio of the price of a certain number of commodities at the present year as against base year.

Quantity Index : As the name suggest, these indices pertain to measuring changes in volumes of commodities like goods produced or goods consumed, etc.

Value Index: These pertain to compare changes in the monetary value of imports, exports, production or consumption of commodities.

14.2.2 PROBLEMS IN THE CONSTRUCTION OF INDEX NUMBERS

The decision regarding the following problems/aspect has to be

taken before starting the actual construction of any type of index numbers.

(i) Purpose of Index numbers under construction
(ii) Selection of base period
(iii) Selection of items
(iv) Selection of source data
(v) Collection of data
(vi) Selection of average
(vii) System of weighting

14.2.3 METHODS OF CONSTRUCTING INDEX NUMBERS

The index number for this purpose is divided into two:

(1) Unweighted Index number
- Simple aggregative
- Simple Average of price relatives

(2) Weighted Index number
- Weighted aggregative
- Weighted Average of price relatives

Unweighted Index number:

There are two methods of constructing unweighted index numbers: (1) Simple Aggregative Method (2) Simple Average of Relative Method

Simple Aggregative Method

In this method, the total price of commodities in a given (current) year is divided by the total price of commodities in a base year and expressed as percentage:

$$\mathbf{P_{01}} = \frac{\sum \mathbf{P1}}{\sum \mathbf{P0}} \times \mathbf{100}$$

Simple Average of Relative Method

In this method, we compute price relatives or link relatives of the given commodities and then use one of the averages such as the arithmetic mean, geometric mean, median, etc. If we use the arithmetic mean as the average, then:

$$\mathbf{P_{01}} = \frac{1}{n} \Sigma \frac{\mathbf{P1}}{\mathbf{P0}} x \mathbf{100}$$

The simple average of relative method is simpler and easier to apply than the simple aggregative method. The only disadvantage is that it gives equal weight to all items.

Example :

The following are the prices of four different commodities for 2017 and 2018. Compute a price index with the (1) simple aggregative

method and (2) average of price relative method by using both the arithmetic mean and geometric mean, taking 2017 as the base.

Commodity	Cotton	Wheat	Rice	Grams
Price in 2017	909	288	767	659
Price in 2018	874	305	910	573

Solution:

Commodity	**Price in 2017** P_0	**Price in 2018** P_1	**Price relative** $P=\frac{P1}{P0}\times 100$	**log p**
Cotton	909	874	69.15	1.9829
Wheat	288	305	105.90	2.0249
Rice	767	910	118.64	2.0742
Grams	659	573	86.95	1.9393
Total	$\Sigma P_0 = 2623$	$\Sigma P_1 = 2662$	**ΣP = 407.64**	**Σ log P = 8.0213**

1. Simple Aggregative Method

   $P_{01} = \frac{\Sigma P1}{\Sigma P0} \times 100 = \frac{2662}{2623} \times 100 =$ **101.49**
2. Simple Average of Price Relative Method (using the arithmetic mean)

   $P_{01} = \frac{1}{n} \Sigma \frac{P1}{P0} x\ 100 = \frac{1}{4}\ (407.64)\ x\ 100 =$ **101.91**

3. Average of price relative method (using the geometric mean)

   $P_{01} = antilog\left(\frac{\Sigma \log P}{4} \right) = antilog\left(\frac{8.0213}{4} \right) =$ **101.23**

Weighted Index number:

When all commodities are not of equal importance, we assign weight to each commodity relative to its importance and the index number computed from these weights is called a weighted index number.

Weighted aggregative index number:

In order to attribute appropriate importance to each of the items used in an aggregate index number some reasonable weights must be used. There are various methods of assigning weights and consequently a large number of formulae for constructing index numbers have been devised of which some of the most important ones are:

1. Laspeyre's Index Number
2. Paasche's Index Number
3. Fisher's Ideal Index Number
4. Marshal-Edge worth Index Number

Laspeyre's Index Number:

In this index number the base year quantities are used as weights, so it also called the base year weighted index.

$$P_{01} = \frac{\sum P1q0}{\sum P0q0} x100$$

Paasche's Index Number:

In this index number the current (given) year quantities are used as weights, so it is also called the current year weighted index.

$$P_{01} = \frac{\sum P1q1}{\sum P0q1} x100$$

Fisher's Ideal Index Number:

The geometric mean of Laspeyre's and Paasche's index numbers is known as Fisher's ideal index number. It is called ideal because it satisfies the time reversal and factor reversal test.

$$P_{01} = \sqrt{\text{Laspeyre's Index} \times \text{Paashe's Index}}$$

$$P_{01} = \sqrt{\frac{\sum P1q0}{\sum P0q0} x \frac{\sum P1q1}{\sum P0q1} x100}$$

Marshal-Edgeworth Index Number:

In this index number the average of the base year and current year quantities are used as weights. This index number was proposed by two English economists, Marshal and Edgeworth.

$$P_{01} = \left(\frac{\sum P1q0 + \sum P1q1}{\Sigma P0q0 + \sum P0q1}\right) x100$$

$$P_{01} = \frac{\Sigma P1(q0 + q1)}{\Sigma P0\ (q0 + q1)} x100$$

Example:

Compute the weighted aggregative price index numbers for 2011 with 2010 as the base year using (1) Laspeyre's Index Number (2) Paasche's Index Number (3) Fisher's Ideal Index Number (4) Marshal-Edgeworth Index Number.

Commodity	**Prices**		**Quantities**	
	2010	**2011**	**2010**	**2011**
A	10	12	20	22
B	8	8	16	18
C	5	6	10	11
D	4	4	7	8

Solution:

Commodity	**Prices**		**Quantities**		P_1q_0	P_0q_0	P_1q_1	P_0q_1
	2010 P_0	**2011** P_1	**2010** q0	**2011** q1				
A	10	12	20	22	240	200	264	220
B	8	8	16	18	128	128	144	144
C	5	6	10	11	60	50	66	55
D	4	4	7	8	28	28	32	32
					$\Sigma P_1q_0 = 456$	$\Sigma P0q0 = 406$	$\Sigma P_1q_1 = 506$	$\Sigma P_0q_1 = 451$

Laspeyre's Index Number:

$$P_{01} = \frac{\sum P1q0}{\sum P0q0} x\ 100 = \frac{456}{406} x\ 100 = \mathbf{112.32}$$

Paasche's Index Number:

$$P_{01} = \frac{\sum P1q1}{\sum P0q1} x\ 100 \quad = \quad \frac{506}{451} x\ 100 = \mathbf{112.20}$$

Fisher's Ideal Index Number

$$P_{01} = \sqrt{\text{Laspeyre's Index} \times \text{Paashe's Index}}$$

$$P_{01} = \sqrt{112.32\ x\ 112.20} \ = \mathbf{112.26}$$

Marshal-Edgeworth Index Number

$$P_{01} = \frac{\Sigma P1(q0 + q1)}{\Sigma P0\ (q0 + q1)} x\ 100 \quad = \quad \left(\frac{456+506}{406+451}\right) x\ 100 \ = \mathbf{112.38}$$

Weighted average of price relatives:

When the specific weights are given for each commodity the weighted index number is calculated by

Weighted Average of Price Relative index = $\frac{\mathbf{\Sigma pw}}{\mathbf{\Sigma w}}$

Where w = the weight of the commodity

p = the price relative index = $\frac{P1}{P0} x\ 100$

When the base year value is P_0q_0 is taken as the weight i.e. w = P_0q_0 then the formula is

Weighted Average of Price Relative index = $\Sigma \frac{\left(\frac{P1}{P0} x\ 100\right) x P0q0}{P0q0} = \frac{\sum P1q0}{\sum P0q0} x\ 100$

This is nothing but Laspeyre's formula

When the weight taken as w = P_0q_1 then the formula is

Weighted Average of Price Relative index = $\Sigma \frac{\left(\frac{P1}{P0} x\ 100\right) x P0q1}{P0q1} = \frac{\sum P1q1}{\Sigma P0q1} x\ 100$

This is nothing but Paasche's formula

Example: Compute the weighted index number for the following data

Commodity	Price		Weight
	Current year	Base year	
X	5	4	40

Y	3	2	60
Z	2	1	20

Solution:

Commodity	Price		Weight	P $=\frac{P1}{P0}x\ 100$	PW
	Current year	Base year			
X	5	4	40	125	5000
Y	3	2	60	150	9000
Z	2	1	20	200	4000
			120		18000

Weighted Average of Price Relative index = $\frac{\Sigma pw}{\Sigma w} = \frac{18000}{120}$ = **150**

14.2.4 QUANTITY OR VOLUME INDEX NUMBER

Price index numbers measures and permit comparison of the price of certain goods; quantity index number, on the other hand, measures the physical volume of production, construction of employment. Though price indices are more widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.

In constructing quantity index numbers, the problems confronting the statistician are analogous to those involved in price indices. We measure changes in quantities, and when we weigh we use prices or values as weights Quantity indices can be obtained easily by changing p to q and q to p in the various formulae discussed above.

Thus, when Laspeyres method is used

$$\mathbf{Q_{01}} = \frac{\sum q1p0}{\sum q0p0} x100$$

When Paasche’s formula is used

$$\mathbf{Q_{01}} = \frac{\sum q1p1}{\sum q0p1} x100$$

When Fisher’s formula is used

$$\mathbf{Q_{01}} = \sqrt{\frac{\sum q1p0}{\sum q0p0} x \frac{\sum q1p1}{\sum q0p1} x100}$$

These formulae represent the quantity index in which the quantities of the different commodities are weighted by their prices.

Example:

Compute the following quantity indices from the data given below (1) Laspeyre's Index Number (2) Paashe's Index Number (3) Fisher's Ideal Index Number.

Commodity	2002		2012	
	Price	Total value	Price	Total value
A	10	200	12	360
B	12	480	15	900
C	15	450	17	680

Solution :

Here instead of quantity, total values are given, hence find quantities of base year and current year

Quantity = $\frac{totalvalue}{price}$

Commodity	p_0	q_0	p_1	q_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	10	20	12	30	200	300	240	360
B	10	40	15	60	400	600	600	900
C	15	30	17	40	450	600	510	680
Total					**1050**	**1500**	**1350**	**1940**

Laspeyre's method $Q_{01} = \frac{\sum q1p0}{\sum q0p0} x\ 100 = \frac{1500}{1050} x\ 100 =$ **142.86**

Paasche's formula $Q_{01} = \frac{\sum q1p1}{\sum q0p1} x\ 100 = \frac{1940}{1350} x\ 100 =$ **143.7**

Fisher's formula $Q_{01} = \sqrt{\frac{\sum q1p0}{\sum q0p0} x \frac{\sum q1p1}{\sum q0p1} x\ 100}$

$= \sqrt{LxP} = \sqrt{142.86\ x143.7}$

= **143.27**

14.2.5 TEST FOR INDEX NUMBER

There are certain tests which are put to verify the consistency, or adequacy of an index number formula from different points of view. The most popular among these are the following tests:

- Order reversal test.
- Time reversal test.
- Factor reversal test.
- Unit test.

At the outset, it should be noted that it is neither possible nor necessary for an index-number formula to satisfy all the tests mentioned above. But, an ideal formula should be such that it satisfies the maximum possible tests which are relevant to the matter under study.

1. Order reversal test:

This test requires that a formula of Index number should be such that the value of the index number remains the same, even if, the order of arrangement of the items is reversed, or altered. As a matter of fact, this test is satisfied by all the formulas of index number explained in this chapter.

2. Time reversal test:

The time reversal test requires that the index number computed backwards should be the reciprocal of the index number computed forward, except for the constant of proportionality

$$P_{01} = \sqrt{\frac{\sum P1q0}{\sum P0q0} x \frac{\sum P1q1}{\sum P0q1}}$$

$$P_{10} = \sqrt{\frac{\sum P0q1}{\sum P1q1} x \frac{\sum P0q0}{\sum P1q0}}$$

$$P_{01} \text{ x } P_{10} = \sqrt{\frac{\sum P1q0}{\sum P0q0} x \frac{\sum P1q1}{\sum P0q1}} x \frac{\sum P0q1}{\sum P1q1} x \frac{\sum P0q0}{\sum P1q0}$$

P_{01} x P_{10} = 1

Laspeyre’s and Paasche‘s method do not satisfy this test but Fisher’s ideal index satisfies this method. Besides both the simple and weighted geometric mean of piece relatives, also, satisfy this time reversal test.

3. Factor reversal test:

This test has also been put forth by Prof. Irving Fisher, in this test the product of price index and quantity index must be equal to the value index. Thus, for the Factor Reversal test, a formula of index number should satisfy the following equation:

Price index × Quantity Index = Value Index

$P_{01} = \sqrt{\frac{\sum P1q0}{\sum P0q0} x \frac{\sum P1q1}{\sum P0q1}}$

$\mathbf{Q_{01}} = \sqrt{\frac{\sum q1p0}{\sum q0p0} x \frac{\sum q1p1}{\sum q0p1}}$

$\therefore$ $\mathbf{P_{01} \text{ x } Q_{01}} = \frac{\sum p1q1}{\sum p0q0}$

Most of the formulae of index number discussed above fail to satisfy this acid test of consistency except that of Prof. Irving Fisher.

4.**Unit test**

This test suggests that the formula for constructing an index should be independent of the unit of measurement in which the prices and quantities are quoted. Except unweighted aggregative index number all other formulas in this chapter satisfy this test.

Example:

Construct Fisher's ideal index for the following data. Test whether it satisfies time reversal test and factor reversal test.

Commodity	Base year		Current year	
	Quantity	Price	Quantity	Price
A	24	20	30	24
B	30	14	40	10
C	10	10	16	18

Solution:

Commodity	q_0	p_0	q_1	p_1	p_0q_0	p_0q_1	p_1q_0	p_1q_1
A	24	20	30	24	480	600	576	720
B	30	14	40	10	420	560	300	400
C	10	10	16	18	100	160	180	288
					1000	**1320**	**1056**	**1408**

Fisher ideal index number $P_{01} = \sqrt{\frac{\sum P1q0}{\sum P0q0} x \frac{\sum P1q1}{\sum P0q1}} x\ 100 = \sqrt{\frac{1056}{1000} x \frac{1408}{1320}} x\ 100$

$= \sqrt{1.056\ x 1.067}$ x 100 $= \sqrt{1.127}$ x 100

$= 1.062 \times 100 \quad = \mathbf{106.2}$

Time Reversal test:

Time Reversal test is satisfied when $P_{01} \times P_{10} = 1$

$$P_{01} = \sqrt{\frac{P1q0}{\sum P0q0} x \frac{\sum P1q1}{\sum P0q1}} = \sqrt{\frac{1056}{1000} x \frac{1408}{1320}}$$

$$P_{10} = \sqrt{\frac{\sum P0q1}{\sum P1q1} x \frac{\sum P0q0}{\sum P1q0}} = \sqrt{\frac{1320}{1408} x \frac{1000}{1056}}$$

$$P_{01} \times P_{10} = \sqrt{\frac{1056}{1000} x \frac{1408}{1320} \text{x} \frac{1320}{1408} x \frac{1000}{1056}} = \sqrt{1} = \mathbf{1}$$

Hence Fisher ideal index satisfy the time reversal test

Factor Reversal Test:

$$P_{01} = \sqrt{\frac{\sum P1q0}{\sum P0q0} x \frac{\sum P1q1}{\sum P0q1}} = \sqrt{\frac{1056}{1000} x \frac{1408}{1320}}$$

$$Q_{01} = \sqrt{\frac{\sum q1p0}{\sum q0p0} x \frac{\sum q1p1}{\sum q0p1}} = \sqrt{\frac{1320}{1000} x \frac{1408}{1056}}$$

$$\therefore \quad P_{01} \times Q_{01} = \sqrt{\frac{1056}{1000} x \frac{1408}{1320} x \frac{1320}{1000} x \frac{1408}{1056}} = \sqrt{\left(\frac{1408}{1000}\right)^2}$$

$$= \frac{1408}{1000} = \frac{\sum p1\, q1}{\sum p0\, q0}$$

Hence Fisher ideal index number satisfy the factor reversal test

14.2.6 CHAIN BASE INDEX NUMBER

In this method, there is no fixed base period; the year immediately preceding the one for which the price index has to be calculated is assumed as the base year. Thus, for the year 1994 the base year would be 1993, for 1993 it would be 1992, for 1992 it would be 1991, and so on. In this way there is no fixed base and it keeps on changing.

The chief advantage of this method is that the price relatives of a year can be compared with the price levels of the immediately preceding year. Businesses mostly interested in comparing this time period rather than comparing rates related to the distant past will utilize this method.

Link relative of current years $= \frac{\text{Price in the Current Year}}{\text{Price in the preceding Year}} x\ 100$

$$\mathbf{Pn - 1\,, n} = \frac{\mathbf{Pn}}{\mathbf{Pn - 1}} \boldsymbol{x}\mathbf{100}$$

Example:

Find the index numbers for the following data taking 2010 as the base year

Years	2004	2005	2006	2007	2008	2009
Price	18	21	25	23	28	30

Solution:

Year	**Price**	**Link Relatives** $\frac{\mathbf{Pn}}{\mathbf{Pn-1}}\boldsymbol{x}\mathbf{100}$	**Chain indices**
2004	18	$\frac{18}{18} x\ 100 = 100$	100
2005	21	$\frac{21}{18} x\ 100 = 116.67$	$\frac{100 \text{ x } 116.67}{100} = 116.7$
2006	25	$\frac{25}{21} x\ 100 = 119.05$	$\frac{116.67 \text{ x } 119.05}{100} = 138.9$
2007	23	$\frac{23}{25} x\ 100 = 92$	$\frac{138.9 \text{ x } 92}{100} = 127.79$
2008	28	$\frac{28}{23} x\ 100 = 121.74$	$\frac{127.79 \text{ x } 121.74}{100} = 155.57$
2009	30	$\frac{30}{28} x\ 100 = 107.14$	$\frac{155.57 \text{ x } 107.14}{100} = 166.68$

CHECK YOUR PROGRESS – 1

1.What is chain base index number
2. What is the formula for Fisher's Ideal Index Number?
3. What is weighted index number?

14.3 COST OF LIVING INDEX NUMBER

Cost of living index numbers measure the changes in the prices paid by consumers for a special "basket" of goods and services during the current year as compared to the base year. The basket of goods and services will contain items like (1) Food (2) Rent (3) Clothing (4) Fuel and Lighting (5) Education (6) Miscellaneous like cleaning, transport, newspapers, etc. Cost of living index numbers are also called consumer price index numbers or retail price index numbers.

14.3.1 CONSTRUCTION OF COST OF LIVING INDEX NUMBERS

The following steps are involved in the construction of Cost of living index numbers.

(1) Class of People:

The first step in the construction of the Cost of living index (CLI) is that the class of people should be defined clearly. It should be decided whether the cost of living index number is being prepared for industrial workers, or middle or lower class salaried people living in a particular area. It is therefore necessary to specify the class of people and locality where they reside.

(2) Family Budget Inquiry:

The next step in the construction of a Cost of living index number is that some families should be selected randomly. These families provide information about the cost of food, clothing, rent, miscellaneous, etc. The inquiry includes questions on family size, income, the quality and quantity of resources consumed and the money spent on them, and the weights are assigned in proportions to the expenditure on different items.

(3) Price Data:

The next step is to collect data on the retail prices of the selected commodities for the current period and the base period when these prices should be obtained from the shops situated in the locality for which the index numbers are prepared.

(4) Selection of Commodities:

The next step is the selection of the commodities to be included. We should select those commodities which are most often used by that class of people.

14.3.2 METHODS TO COMPUTE COST OF LIVING INDEX NUMBERS

There are two methods to compute cost of living index numbers: (1) Aggregate Expenditure Method (2) Family Budget Method

Aggregate Expenditure Method

In this method, the quantities of commodities consumed by the particular group in the base year are estimated and these figures or their proportions are used as weights. Then the total expenditure of each commodity for each year is calculated. The price of the current year is multiplied by the quantity or weight of the base year. These products are added. Similarly, for the base year the total expenditure of each commodity is calculated by multiplying the quantity consumed by its price in the base year. These products are also added. The total expenditure of the current year is divided by the total expenditure of the base year and the resulting figure is multiplied by 100 to get the required index numbers. In this method, the current period quantities are not used as weights because these quantities change from year to year.

$$P_{01} = \frac{\sum P1q0}{\sum P0q0} x100$$

Here,
P_1 - Represent the price of the current year,
P_0 - Represents the price of the base year and
q_0- Represents the quantities consumed in the base year.

Family Budget Method:

In this method, the family budgets of a large number of people are carefully studied and the aggregate expenditure of the average family for various items is estimated. These values are used as weights. The current year's prices are converted into price relatives on the basis of the base year's prices, and these price relatives are multiplied by the respective values of the commodities in the base year. The total of these products is divided by the sum of the weights and the resulting figure is the required index numbers.

$$P_{01} = \frac{\sum WI}{\sum W} x100$$

Here, $I=\frac{\sum P1}{\sum P0} x100$ and $\Sigma W=P_0q_0$

Example:

Construct the cost of living index number for 2018 on the basis of 2017 from the following data using (1) Aggregate Expenditure Method (2) Family Budget Method.

Commodity	Quantity Consumed in 2017 (in quintal)	Prices	
		2017	2018
A	6	315.75	316.00
B	6	305.00	308.00
C	1	416.00	419.00
D	6	528.00	610.00
E	4	120.00	119.50
F	1	1020.00	1015.00

Solution:

The cost of living index number of 2018 by Aggregate Expenditure method:

Commodity	**Quantity Consumed in 2017 (in quintal) q_0**	**Prices**		**P_1q_0**	**P_0q_0**
		2017 P_0	2018 P_1		
A	6	315.75	316.00	1896	1894.50
B	6	305.00	308.00	1848	1830.00
C	1	416.00	419.00	419	416.00
D	6	528.00	610.00	3660	3168.00
E	4	120.00	119.50	478	480.00
F	1	1020.00	1015.00	1015	1020.00
				ΣP_1q_0	**ΣP_0q_0 =**

				= 9316	**8808.5**

The cost of living index number of 2018 is

$$\mathbf{P_{01}} = \frac{\sum P1q0}{\sum P0q0} x100 = \frac{9316}{8808.5} x100 = \mathbf{105.76}$$

The cost of living index number of 2018 by Family Budget Method:

Commodity	Quantity Consumed in 2017 (in quintal) q_0	Prices		$W=P_0q_0$	$I=\frac{\sum P1}{\sum P0}x100$	Product WI
		2017 P_0	2018 P_1			
A	6	315.75	316.00	1894.5	100.08	189601.56
B	6	305.00	308.00	1830.0	100.98	184793.40
C	1	416.00	419.00	416.0	100.72	41899.52
D	6	528.00	610.00	3168.0	115.53	365999.04
E	4	120.00	119.50	480.0	99.58	47798.4
F	1	1020.00	1015.00	1020.0	99.51	101500.20
				ΣW = 8808.5		ΣWI =931592.12

The cost of living index number of 2018 is

$$\mathbf{P_{01}} = \frac{\sum WI}{\sum W} x\ 100 = \frac{931592.12}{8808.5} = \mathbf{105.76}$$

14.3.3 USES OF COST OF LIVING INDEX NUMBER

- They indicate the changes in the consumer prices. Thus they help government in formulating policies regarding control of price, taxation, imports and exports of commodities, etc.
- They are used in granting allowances and other facilities to employees
- They are used for the evaluation of purchasing power of money. They are used for deflating money
- They are used for comparing changes in the cost of living of differenc classes of people

14.4 USES OF INDEX NUMBER

The main uses of index numbers are given below.

- Index numbers are used in the fields of commerce, meteorology, labour, industry, etc.
- Index numbers measure fluctuations during intervals of time, group differences of geographical position of degree, etc.
- They are used to compare the total variations in the prices of different commodities in which the unit of measurements differs with time and price, etc.
- They measure the purchasing power of money.
- They are helpful in forecasting future economic trends.
- They are used in studying the difference between the comparable categories of animals, people or items.
- Index numbers of industrial production are used to measure the changes in the level of industrial production in the country.
- Index numbers of import prices and export prices are used to measure the changes in the trade of a country.

Index numbers are used to measure seasonal variations and cyclical variations in a time series.

14.5 LIMITATIONS OF INDEX NUMBER

- They are simply rough indications of the relative changes.
- The choice of representative commodities may lead to fallacious conclusions as they are based on samples.
- There may be errors in the choice of base periods or weights, etc.
- Comparisons of changes in variables over long periods are not reliable.
- They may be useful for one purpose but not for another.
- They are specialized types of averages and hence are subject to all those limitations which an average suffers from.

CHECK YOUR PROGESS - 2

4. What are the methods to compute Cost of Living Index numbers?

5. What are thepopularTests for Index number?
6.Write a few uses of index number.

14.6 SUMMARY

- Index numbers are meant to study changes in the effects of factors which cannot be measured directly. According to Bowley, “Index numbers are used to measure the changes in some quantity which we cannot observe directly”.

- . Price Index Quantity Index Value Index.Quantity Index Numbers are the types of index numbers.
- Price index numbers measures and permit comparison of the price of certain goods; quantity index number, on the other hand, measures the physical volume of production, construction of employment. Though price indices are more widely used, production indices are highly significant as indicators of the level of output in the economy or in parts of it.
- There are certain tests which are put to verify the consistency, or adequacy of an index number formula from different points of view. The most popular among these are the following tests: (1)Order reversal test.(2) Time reversal test. (3) Factor reversal test. (4)Unit test.
- In this method, there is no fixed base period; the year immediately preceding the one for which the price index has to be calculated is assumed as the base year.
- Cost of living index numbers measure the changes in the prices paid by consumers for a special "basket" of goods and services during the current year as compared to the base year.
- There are two methods to compute cost of living index numbers: (1) Aggregate Expenditure Method (2) Family Budget Method.

8.7 KEY WORDS

Index numbers,Price Index, Quantity Index, Value Index, Laspeyre's Index Number, Paasche's Index Number, Fisher's Ideal Index Number, Marshal-Edge worth Index Number,Order reversal test, Time reversal test, Factor reversal test, Unit test,Chain Base index number, Cost of living index number.

8.8 ANSWERS TO CHECK YOUR PROGRESS

1. In this method, there is no fixed base period; the year immediately preceding the one for which the price index has to be calculated is assumed as the base year.

2.$P_{01} = \sqrt{\text{Laspeyre's Index} \times \text{Paashe's Index}}$

$$P_{01} = \sqrt{\frac{\sum P1q0}{\sum P0q0} x \frac{\sum P1q1}{\sum P0q1} x100}$$

3.When all commodities are not of equal importance, we assign weight to each commodity relative to its importance and the index number computed from these weights is called a weighted index number.

4.There are two methods to compute cost of living index numbers: (1)

Aggregate Expenditure Method (2) Family Budget Method

5. Order reversal test, Time reversal test, Factor reversal test, Unit test

6. Index numbers are used in the fields of commerce, meteorology, labour, industry, etc.Index numbers measure fluctuations during intervals of time, group differences of geographical position of degree, etc.They are used to compare the total variations in the prices of different commodities in which the unit of measurements differs with time and price, etc.They measure the purchasing power of money.They are helpful in forecasting future economic trends

14.9 QUESTIONS AND EXERCISE

SHORT ANSWER QUESTIONS

1. Define index number and write the uses of index numbers
2. State the types of index numbers
3. State the methods of constructing consumer price index

LONG ANSWER QUESTIONS

1. Compute (1) Laspeyre's (2) Paasche's index number for the 2001 from the following

Commodity	Price		Quantity	
	2002	2010	2002	2010
W	4	6	8	7
X	3	5	10	8
Y	2	4	14	12
Z	5	7	19	11

2. Calculate Fisher's ideal index method for the following data

Commodity	2011		2012	
	Price	Quantity	Price	Quantity
A	2	7	3	5
B	5	11	6	10
C	3	14	5	11
D	4	16	4	18

3. Construct the consumer price index number of 2015 on the from the following data using

(i) the average expenditure method and

(ii) the family budget method

Commodity	Quantity Consumed in 2014	Prices	
		2014	2015
A	6 Kg	5	7
B	6 Quintal	6	6
C	5 Quintal	5	4
D	6 Quintal	7	7
E	4 Quintal	8	8
F	5 Kg	9	9

14.10 FURTHER READINGS

(1) Statistics (Theory & Practice) by Dr. B.N. Gupta. SahityaBhawan Publishers andDistributors (P) Ltd., Agra.
(2) Statistics for Management by G.C. Beri. Tata McGraw Hills Publishing CompanyLtd., New Delhi.
(3) Business Statistics by Amir D. Aczel and J. Sounderpandian. Tata McGraw HillPublishing Company Ltd., New Delhi.
(4) Statistics for Business and Economics by R.P. Hooda. MacMillan India Ltd., NewDelhi.
(5) Business Statistics by S.P. Gupta and M.P. Gupta. Sultan Chand and Sons.,NewDelhi.
(6) Statistical Method by S.P. Gupta. Sultan Chand and Sons., New Delhi.

DISTANCE EDUCATION – CBCS – (2018 – 2019 Academic Year Onwards)

Question Paper Pattern – BUSNESS STATISTICS

(UG Programs)

Time: 3 Hours **Maximum: 75 Marks**

Part – A (10 x 2 = 20 Marks)

Answer all questions

1. What is chi-square test?
2. What is meant by analysis of variance?
3. Write short note on index number.
4. Describe Type II error
5. Write any four advantages of statistics
6. Explain the term Probable Error
7. What is meant by Binomial Distribution?
8. Define the term “Correlation”
9. What is meant by Regression?
10. What is forecasting?

Part – B (5 x 5 = 25 Marks)

Answer all questions choosing either (A) or (B)

11. (A) what are the importances of statistics?

(or)

(B) Calculate median from the following data

Profit :	100 – 200	200 – 300	300 – 400	400 – 500	500 – 600	600 – 700	700 – 800
No. of Shops:	10	18	20	26	30	28	18

12. A) Find rank correlation.

X :	1	2	3	4	5
Y :	5	4	3	4	1

(or)

(B) What is the probability that a leap year will contain 53 Sundays?

13. (A) The manufacture of a certain make of electric bulbs claims that the bulbs have a mean life of 25 months with a life of 25 onths with a standard deviation of 5 months. A random sample of 6 such bulbsgave the following values.

Life of months: 24 26 30 20 20 18

Can you regard the producer’s claim to be valid at 1% level of significance? V = 4.032

(or)

(B). Find standard deviation from the following observations.

Size	: 120	125	130	135	140	145	150	155	160
Frequencies	: 2	3	3	1	2	7	4	2	8

14. (A) The first of two samples consists of 23 pairs and gives a correlation coefficient of 0.5 while the second of 28 pairs has the correlation coefficient of 0.8. Are these values significantly different?

(or)

(B) Explain in brief the assumptions of ANOVA.

15. (A) Calculate the 3 yearly and 5 yearly moving averages of the data

Years	1	2	3	4	5	6	7	8	9	10	11	12
sales	5.2	4.9	5.5	4.9	5.2	5.7	5.4	5.8	5.9	6.00	5.2	4.8

(or)

(B) Calculate trend values by the method of least square from data given below and estimate the sales for 2003

Years:	1996	1997	1998	1999	2000
Sales of Co.A, (₹ Lakhs)	70	74	80	86	90

Part – C (3 x 10 = 30 Marks)

Answer any three out of five questions

16. Eight coins are tossed simultaneously 256 times. Number of heads observed at each throw is recorded as given below:

No. of Heads:	0	1	2	3	4	5	6	7	8
Frequencies :	2	6	30	52	67	56	32	10	1

Fit a binomial distribution and the expected frequency. Also find the mean and standard deviation

17. The quantity of raw materials purchased by a company at the specified prices during the 12 months of 2017 is given as follows:

Months	J	F	M	A	M	J	JY	A	S	O	N	DEC
Price in Rs	96	110	100	90	86	92	112	112	108	116	86	92
Quantity Per kg	250	200	250	280	300	300	220	220	200	210	300	250

a. Find the regression equations based on the above data.

b. Can you estimate the approximate quantity likely to be purchased if the price shoots up to Rs.124 per kg?

18. 200 digits are chosen at random from a set of tables. The frequencies of the digits are as follows

Digit :	0	1	2	3	4	5	6	7	8	9
F :	18	19	23	21	16	25	22	20	21	15

Use Chi-Square test to assess the correctness of the hypothesis that the digits were distributed in equal numbers in the tables from which they were chose. V, 9 = 16.22

19. The sales data of an item in six shops before and after a special promotional campaign are as under:

Before Campaign:	53	28	31	48	50	42
After Campaign :	58	29	30	55	56	45

Can the campaign be judged to be a success? Test at 5% level of significance. V= t0.05 = 2.57

20. Construct L & P method of index.

Types	Price		Quantity	
	2012	2014	2012	2014
X	9	15	5	5
Y	4	12	10	11
Z	1	5	6	6